REGULATION AND EVOLUTIONARY ORIGINS OF REPULSIVE GUIDANCE MOLECULE C / HEMOJUVELIN EXPRESSION: A MUSCLE-ENRICHED GENE INVOLVED IN IRON METABOLISM

by

Christopher John Severyn

B.A. (University of California, Berkeley) 2002

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

BIOCHEMISTRY AND MOLECULAR BIOLOGY

Presented to the Department of Biochemistry & Molecular Biology

OREGON HEALTH & SCIENCE UNIVERSITY

School of Medicine

July 2010

Department of Biochemistry & Molecular Biology,
School of Medicine
OREGON HEALTH & SCIENCE UNIVERSITY

CERTIFICATE OF APPROVAL

This is to certify that the Ph.D. dissertation of

Christopher J. Severyn

has been approved by the following:

Peter Rotwein, M.D., dissertation advisor
Professor of Biochemistry and Department Chair

Maureen Hoatlin, Ph.D., committee chair
Associate Professor of Biochemistry

Matt Thayer, Ph.D., committee member
Professor of Biochemistry

Ujwal Shinde, Ph.D., committee member
Professor of Biochemistry

Jim Lundblad, M.D., Ph.D., committee member
Associate Professor of Medicine and Chief of Endocrinology

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

α	alpha
A (amino acid)	alanine
Å	Angstrom
Abs	absorption
Ad	adenovirus
AMP	adenosine monophosphate
Arg	arginine
Asn	asparagine
Asp	aspartic acid
ATP	adenosine triphosphate
β	beta
BAC	bacterial artificial chromosome
BCA	bicinchoninic acid
bHLH	basic helix-loop-helix
BMP	bone morphogenetic protein
bZIP	basic leucine zipper domain
C (amino acid)	cysteine
C-	carboxyl-
°C	degrees Celsius
Ca	calcium
cAMP	cyclic adenosine monophosphate
CD	circular dichroism
CDS	(protein) coding sequence in mRNA
cDNA	complementary DNA
CHAPS	3-[(3-cholamidopropyl)dimethyl-ammonio]-1-propanesulfonate
Chd	chromodomain helicase
ChIP	chromatin immunoprecipitation
ChIP-Seq	chromatin immunoprecipitation followed by DNA sequencing

CREB	cyclic-AMP response element binding protein
Cl	chlorine
cm	centimeter (10^{-3} meter)
CNE	conserved non-coding element(s)
Cu^{2+}	copper
CY3	carboxymethylindocyanine dye 3.18
CY5	carboxymethylindocyanine dye 5.18
Cys	cysteine
δ	delta
Δ	delta, or the difference between two values
D (amino acid)	aspartic acid
Da	Dalton, unit of atomic mass
ΔG	change in Gibbs free energy
ΔH	change in enthalpy
ΔS	change in entropy
DM	differentiation media
DMSO	dimethylsulfoxide
DMEM	Dulbecco's modified Eagle's medium
DRB	5,6-dichloro-1-β-D-ribofuranosylbenzimidazole
DRG	dorsal root ganglion
DNA	deoxyribonucleic acid
DTT	dithiothreitol
ε	epsilon
ε (absorbance)	extinction coefficient
E (amino acid)	glutamic acid
E	embryonic day
EDTA	ethylenediamine-tetraacetic acid
e.g.	*exempli gratia* (Latin), "for example"
eIF	eukaryotic initiation factor (e.g., eIF4)
EM	electron microscopy
EMSA	electrophoretic mobility shift assay

EST	expressed sequence tag
ETS	E26 avian retrovirus Transformation-Specific
et al.	*et alii* (Latin), “and others”
F (amino acid)	phenylalanine
FBS	fetal bovine serum
γ	gamma
g (mass)	gram
G (amino acid)	glycine
GDP	guanosine diphosphate
Glu	glutamic acid
GPI	glycosylphosphatidylinositol
GTP	guanosine triphosphate
hr	hour(s)
H (atom)	hydrogen
H (amino acid)	histidine
HIV	human immunodeficiency virus
HJV	Hemojuvelin
HNF	hepatocyte nuclear factor
I (amino acid)	isoleucine
i.e.	*id est* (Latin), “that is”
IGF	insulin-like growth factor
IPTG	isopropyl β-thiogalactoside
IR	infrared
IRE	iron response element
IRES	internal ribosome entry site
IRP	IRE (iron response element) binding protein
ItAF	IRES (internal ribosome entry site) *trans*-Acting Factor
JH	juvenile hemochromatosis
κ	kappa
k	rate constant
K (atom)	potassium

K (amino acid)	lysine
kDa	kilodalton (10^3 Daltons)
kcal	kilocalories (10^3 calories)
L (amino acid)	leucine
Lix1	limb expression 1
Lys	lysine
M	molar
M (amino acid)	methionine
MADS	MCM1–agamous–deficiens–serum response factor
max	maximum
MCM1	minichromosome maintenance
MCK	creatine kinase-muscle
Mctp2	multiple C2 domains, transmembrane 2
MEF2	myocyte enhancer factor 2
Met	methionine
Mg	magnesium
MHC	myosin heavy chain (myh3)
µ	micro- = $1x10^{-6}$
µg	microgram (10^{-6} gram)
mg	milligram (10^{-3} gram)
min	minute
miRNA	micro RNA
µl	microliter (10^{-6} liter)
ml	milliliter (10^{-3} liter)
µM	micromolar (10^{-6} Molar)
mM	millimolar (10^{-3} Molar)
mol	mole ($6.02x10^{23}$ molecules)
MOPS	3-(*N*-morpholino)-propanesulfonic acid
µs	microsecond (10^{-6} second)
ms	millisecond (10^{-3} second)
MW	molecular weight

MyoD	myogenic differentiation-1
myog.	myogenin
n	refractive index for optics
n-	nano- = $1x10^{-9}$
N (amino acid)	asparagine
N (atom)	nitrogen
N-	amino-
Na	sodium
NCS	newborn calf serum
NF-κB	nuclear factor kappa-light-chain-enhancer of activated B cells
nm	nanometer (10^{-9} meter)
NMR	nuclear magnetic resonance
NR	nuclear receptor (transcription factor) family
O (atom)	oxygen
OD	optical density
Ω	omega
P (amino acid)	proline
P (atom)	phosphate
PCR	polymerase chain reaction
pH	potential of hydrogen ($-\log_{10}[H]$)
Phe	phenylalanine
PI-PLC	phosphoinositide-specific phospholipase C
pKa	acid dissociation constant
pmol	picomole (10^{-12} mole)
Polr3gl	polymerase (RNA) III (DNA directed) polypeptide G-like
PPC	pro-protein convertase
PTB	polypryrimidine tract binding protein
ρ	Rho (being used for density)
Q (amino acid)	glutamine
R (amino acid)	arginine
RACE	rapid amplification of cDNA ends

RGD (motif)	arginine-glycine-aspartic acid
RGM	Repulsive Guidance Molecule
RLU	relative luciferase units
RNA	ribonucleic acid
RT-PCR	reverse transcription polymerase chain reaction
s	second(s)
S (amino acid)	serine
S (atom)	sulfur
SDS	sodium dodecyl sulfate
SDS PAGE	SDS polyacrylamide gel electrophoresis
SEM	standard error of the mean
Ser	serine
Slco/SLCO	solute carrier organic anion transporter family
St8sia/ST8SIA	ST8 α-N-acetyl-neuraminide α-2,8-sialyltransferase
STAT	signal transducers and activators of transcription
T (amino acid)	threonine
τ	Tau (being used for decay lifetime)
t1/2	half-life
TFBS	transcription factor binding sites
TGF	transforming growth factor
TK	thymidine kinase (referring to the TK promoter)
Tm	melting temperature
TCA	trichloroacetic acid
TM	transmembrane
Tris	2-amino-2-(hydroxymethyl)-1,3-propanediol
Trp	tryptophan
TSS	transcription start site
Txnip	thioredoxin interacting protein
uORF	upstream open reading frame
UTR	untranslated region
Unc	uncoordinated (gene)

UV	ultraviolet
V (amino acid)	valine
vWD	von Willebrand type D
W (amino acid)	tryptophan
WT	wild type
Y (amino acid)	tyrosine
ZAN	zonadhesin

ACKNOWLEDGEMENTS

I first want to thank Peter Rotwein for his guidance, advice, and ability to challenge me in all aspects of my scientific life in order to provide the foundation for my growth as a physician-scientist. I want to extend my gratitude and appreciation for all the help, discussions, and recommendations from the members of my research advisory committee— Maureen Hoatlin, Ujwal Shinde, Matt Thayer, and Jim Lundblad. Collectively and individually, you all have taught me more than you may possibly realize. Thank you.

I also want to extend a thank you to the members of the Rotwein Laboratory, both past and present, for providing me with the environment and opportunity to learn about a variety of aspects of a social and productive lab. In particular, I want express my appreciation to Lisa Wilson, David Kuninger, Mahta Nili, and Aditi Mukherjee, for all your help and patience with my incessant questions, no matter how trivial, and of our discussions ranging from science to politics and current affairs to knowing (and debating about) who the Greatest University in the Pac-10 (or another conference) might actually be. Thank you for making the time in and out of the lab truly wonderful. To Mona Hwang, Michael Thelen, Rod Balhorn, and Shelly Corzett at LLNL, for providing me the opportunity as a freshman at Cal the ability to test my hands at fundamental sciences, and allow my love for biology to grow during that time. You helped provide the foundation for my excitement of biology, biochemistry, and biophysics inside the laboratory.

I would be amiss if I forgot to thank my closest friends, Ellena Mar, Jesse Koh, and Jill Wentzell, for their support, encouragement, laughter, and making my days that much brighter. I look forward to the adventures that unfold together over the years… in due time, of course.

Finally, I want to thank the Severyn, Louie, Palma, Askins, and Paulsen families. Without you all, I would not be the person who I am today.

This work is dedicated…

To my family, whose love, support, countless sacrifices, encouragement, and guidance over the years helped me to attend and learn at some of the greatest Universities in the world.

To Oregon for allowing me to explore the numerous outdoor offerings, enjoy the warmth of the people of the Pacific Northwest, as well as the opportunity to study and appreciate the wonders of molecular biology and medicine.

And to California. *Fiat Lux*.

PEER-REVIEWED PUBLICATIONS ARISING FROM THIS DISSERTATION

Severyn, C. J., Shinde, U., and Rotwein, P. (2009) "Molecular biology, genetics and biochemistry of the repulsive guidance molecule family." *Biochem J.*, **v.422**, 393-403

Severyn, C. J., and Rotwein, P. (2010) "Proximal promoter elements control repulsive guidance molecule c / hemojuvelin gene transcription in skeletal muscle." *Submitted to Genomics*

Severyn, C. J., and Rotwein, P. "Repulsive Guidance Molecule c / Hemojuvelin regulation by a post-transcriptional element in the 5'-untranslated region." *In preparation, June 2010.*

ABSTRACT

Ominum rerum principia parva sunt.

"Everything has a small beginning…." –Cicero, 'De finibus.'

The mechanisms that regulate gene expression are complex, however many features have been well-conserved through evolution. Repulsive guidance molecule c (RGMc), or hemojuvelin (HJV), is a member of a three gene family that in most vertebrates plays a critical role in iron metabolism, yet virtually nothing is known about the regulation of RGMc gene expression. To better understand the mechanisms that regulate RGMc expression, this dissertation investigates the molecular biology and biochemistry of the RGM family, presents the first detailed analysis of the RGMc promoter and data to support a post-transcriptional mechanism for RGMc gene regulation, and integrates the findings into an understanding of the molecular evolution of the RGM family of genes.

This dissertation discusses three main topics: (1) the genomic structure of the RGM family of genes, (2) the mechanisms of transcriptional and post-transcriptional regulation focusing on RGMc, and (3) the molecular evolution of the RGM family. *The long-range and overarching goal of this dissertation* is to enhance understanding of the molecular mechanisms by which evolution has shaped the regulation of gene expression in a family of genes, like the RGM family.

Following a brief introductory chapter on gene regulation, this work addresses the molecular biology and biochemistry of the RGM family, beginning with the structures of the genomic loci and organization of the genes across multiple organisms. In addition, chapter 2 attempts to define and critically evaluate what is known about the RGM family, and identify critical gaps in our understanding of the gene family. The molecular evolution of the gene family is presented along with the first *ab intio* structural model to permit future work for investigators in the field.

Chapter 3 of this thesis focuses on defining the detailed gene struture of RGMc, its transcripts, and mechanisms of transcriptional regulation of the gene. Using reporter gene experiments, three critical regions of the proximal promoter are identified that are responsible for RGMc transcriptional activation in skeletal muscle, comprising paired E-boxes, a putative Stat and/or Ets element, and a MEF2 site. In non-muscle cells, expression of the muscle transcription factors myogenin and MEF2C can stimulate RGMc promoter function suggesting that these factors are important components for the expression of the gene in skeletal muscle. As these elements are highly conserved in RGMc from multiple mammalian species, the results presented in chapter 3, coupled with the evolutionary analysis in chapter 2, support the hypothesis that RGMc has been a muscle-enriched gene throughout its evolutionary history.

Finally, the 4th chapter reveals a novel region, called the ε-element, in the untranslated region of the RGMc transcript that operates via a post-transcriptional mechanism. RT-PCR results demonstrate a constant steady-state level of mRNA whether this element is present or absent, but an order of magnitude increase in reporter expression only when the element is present. Additional data reveals that RGMc is not regulated by iron levels prior to the formation of nascent protein. These data suggest that the ε-element controlling RGMc expression may be an example of a small, but growing number of regulatory mechanisms that utilize the 5'-untranslated region (UTR) to enhance translation of specific mRNA transcripts into a nascent protein.

In summary, this dissertation reveals the promoter of RGMc (the first example in the entire three gene RGM family), a possible positive translational control element in the 5'UTR of RGMc, the first *ab initio* structure of the RGM protein family, and provides a foundation to understand the molecular evolution of RGMc, and how this knowledge may be applicable to gaining insight into the structure, function, and development of gene families and their tissue-specific patterns of expression.

Key words: repulsive guidance molecule (RGM), RGMc, Hemojuvelin, axon guidance, gene evolution, gene structure, iron metabolism, protein modeling, transcriptional regulation, post-transcriptional regulation, translational regulation, molecular evolution, MEF2C, myogenin, 5'UTR, HNF4 alpha, STAT, ETS, bone morphogenetic protein, BMP, neogenin, pediatric diseases, juvenile hemochromatosis

Chapter 1

Introduction

"The most erroneous stories are those we think we know best -- and therefore never scrutinize or question." –Stephen Jay Gould

"Dubitando quippe ad inquisitionem venimus; inquirendo veritatem percipimus," (by doubt indeed we come to questioning; by questioning, we perceive the truth). –From the Prologue to Peter Abélard's '*Sic et Non*.'

1.1: General Overview.

Iron-related metabolic and hematologic disorders affect millions of individuals worldwide. Repulsive guidance molecule c (RGMc), or hemojuvelin (HJV), is a gene shown to be critical for iron regulation [1-3], as inactivating mutations in RGMc/HJV cause juvenile hemochromatosis, a severe systemic iron overload disorder in humans [3]. The molecular mechanisms responsible for the regulation of RGMc under physiological and pathological conditions, as well as many of the most fundamental aspects of the biology of RGMc are unknown. In addition, understanding the regulation of RGMc expression has been complicated by its unique tissue distribution, being expressed exclusively in striated muscle and liver.

The primary focus of this work will be (i) to define the structure of the RGM family of genes from the genomic level to an *ab* initio protein model, (ii) understand the mechanisms responsible for the regulation of RGMc transcription in skeletal muscle, and (iii) gain insight into the post-transcriptional regulatory mechanisms that may control RGMc expression in multiple tissues. *The overarching theme of this dissertation* is to enhance understanding of the molecular mechanisms by which evolution has shaped the regulation of gene expression in a family of genes, like the RGM family. From the experimental data, this dissertation will discuss the evolutionary implications of these regulatory mechanisms and provide a foundation for future experimentation within the RGM family of genes, as well as implications for how novel expression patterns may arise from conserved regulatory mechanisms.

The first part of this introductory chapter will briefly explore the regulatory mechanisms of gene expression from an evolutionary perspective, focusing on control at the transcriptional and post-transcriptional levels. Next, there will be a brief overview of

where RGMc appears to be a critical regulator of systemic iron metabolism. Finally, I will discuss the overall objectives of each chapter of this dissertation.

1.2: Expression and Regulation of RGMc

The control of gene expression in complex tissues such as skeletal muscle and the liver involves an elaborate interplay between signaling pathways and their downstream effectors, which include factors that affect transcription, mRNA stability, translation of the protein, and post-translational processes on the protein. How tissues specifically control the initial expression of tissue-restricted genes and in response to stimuli remains an ongoing challenge. RGMc is a gene whose expression is restricted to striated muscle and liver, yet its loss or mis-expression has a profound impact on systemic iron metabolism. The signaling pathways and fundamental regulatory mechanisms that control RGMc gene expression in development and normal physiology remain largely unknown. To date there are no other genes whose expression is restricted to striated muscle and liver, making RGMc unique in that understanding its regulatory mechanisms will provide insight into systemic iron regulation and tissue-specific gene expression. ***The main focus of this work*** will be to define the mechanisms responsible for the regulation of RGMc transcription in skeletal muscle, and the post-transcriptional mechanisms in skeletal muscle and liver. The *long range goals* are to understand the mechanisms of tissue-specific regulation of RGMc/HJV in response to developmental or physiological cues and how this expression influences iron metabolism in normal and diseased states. For example, this work could be applied towards a tissue-specific knockout of RGMc to determine the possible functions of the gene in each tissue in which it is expressed. To place these studies in the appropriate context, outlined below is the current understanding

of the regulatory mechanisms of gene expression from an evolutionary perspective, as well as the role of RGMc in systemic iron homeostasis.

1.3: Gene Regulation: Transcription from an evolutionary perspective.

The first example of gene control began with the experiments by François Jacob and Jacques Monod in the 1960's. Using biochemistry and genetics on the bacterial lactose operon (set of genes), they demonstrated the primary recognition of (i) protein-binding to sequences on genes and (ii) that this binding leads to the regulation of transcription [4]. Several elegant experiments over the ensuing decades performed by teams of researchers using both prokaryotes and eukaryotes followed these groundbreaking discoveries and led to an ever increasing understanding of the complexity of transcriptional regulation. Despite this complexity, the importance of a sequence-specific DNA-binding and subsequent recruitment of large numbers of enzymes and structural proteins that allow the RNA polymerase to transcribe the nascent messenger RNA (mRNA) [5] can be distilled into a simple premise: the mechanisms of regulated recruitment that permit transcription are highly conserved [6-9]. This is especially true in eukaryotes, where large numbers of proteins (estimated to be at least 60-80 individual subunits [5, 7, 8, 10, 11]) regulate the initiation of transcription. Nevertheless, the fundamental concept of 'regulated recruitment' remains a central theme [8, 9]. Classic examples are the yeast Gal4 system [6] and the yeast two-hybrid system [4, 12] where the transcriptional activation domain and DNA-binding domains are separable. This makes intuitive sense from an evolutionary perspective, as recruitment is an easier way for systems to evolve, especially for complex signaling with multiple inputs [7-9]. When biological events are regulated by recruitment, there is almost always a background level, and the system is simply increasing the probability with which a spontaneous event might occur, thus

increasing the level of expression in the case of gene regulation [7, 8]. From an evolutionary perspective, 'regulated recruitment' is an ideal way for systems to evolve complex regulatory mechanisms to control gene expression.

As early as 1975, King and Wilson suggested during the course of evolution, subtle changes in the regulation of genes following gene duplication was the most parsimonius explanation for appearance of new gene families with different functions (and/or patterns of expression) [13]. As much of the regulatory machinery is well-conserved in eukaryotes [14-17], it is the regulated recruitment of various transcription factors that likely controls the restricted patterns of expression seen in a large number of gene families, which almost always involves changes to the transcription factor binding sites [14, 17, 18]. This fundamental concept of transcription factor binding site conservation, coupled with the large explosion of genomic sequences, has led many to propose that gene families of recently duplicated, but subsequently diverged genes may provide a unique opportunity for comparative analysis of regulatory elements (a concept called 'phylogenetic footprinting') [19-21]. One of the major challenges with phylogenetic footpriniting is the ability to obtain reliably orthologous (similar or identical genes following a speciation event) promoter regions, a process that is actually quite difficult due to the numerous idiosyncrasies of vertebrate genomics [20, 21]. However, this challenge is greatly reduced when there are high quality functional data about one of the promoter regions being compared. In this dissertation, the promoter region of mouse RGMc is characterized (chapter 3), and subsequently compared with the genomic sequences of RGMc from other organisms. By having a data set that is functionally defined, and comparing it to genomic sequence data, inferences about the evolution of RGMc gene regulation, and potentially the entire RGM family, can be made. As will be discussed in the following chapters, the fundamental findings presented in this dissertation may provide a foundation for future regulatory studies for the entire RGM

family, all based on the premise that evolution has already performed the experiment on whole organisms.

1.4: Gene Regulation: Translation.

Often gene regulation occurs at the level of transcription intitation, one of the earliest steps in a biochemical pathway [5, 11, 22]. In addition to transcriptional regulatory mechanisms, gene expression can be controlled at multiple levels including post-transcriptional processes like translation of the mRNA. Both transcription and translation are critical for the cell, as both of these biosynthetic steps require a large investment of energy [22] and mis-regulation can have deleterious consequences [5, 23]. There are several compelling reasons for a cell to use translational control in its arsenal of regulatory mechanisms including (i) rapidity of need for certain proteins, (ii) finer control, including spatial needs of the cell, (iii) regulation of large genes, (iv) in systems that lack transcriptional control (e.g., reticulocytes, oocytes, RNA viruses), and several other possibilities as reviewed in Mathews, et al., [22]. Chapter 4 of this thesis will investigate an element in the untranslated region of RGMc mRNA that appears to be regulated at the level of translation.

1.5: Iron Metabolism in Eukaryotes.

Iron is an essential element critical for numerous cellular processes and will be reviewed in the context of RGMc in chapters 2 and 3. While the primary focus of this work is to understand the regulatory processes that control RGMc gene expression, it is important to consider the known phenotype when RGMc / HJV is mutated, that is severe iron overload. In 2004, RGMc / HJV was found to be an important component of whole body

iron metabolism, and RGMc was suggested to regulate the peptide hormone hepcidin. Figure 1.1 outlines the current understanding of RGMc's relationship to iron regulation, along with the implications of muscle being added to the iron homeostatic pathways (as RGMc is expressed in striated muscle and liver). As will be noted in chapter 2, this model is oversimplified as RGMc appears to be processed into at least four distinct isoforms capable of binding different proteins with variable affinities (see Fig. 2.5 for details). Nevertheless, the clinical phenotype of patients with mutations in human HJV (RGMc / HFE2) is that of decreased levels of hepcidin in the blood and the urine, and subsequent iron overload. Experiments presented in chapter 4 support the hypothesis that any effects of iron on RGMc expression are unlikely to occur prior to the appearance of protein. The iron-sensing mechanism for RGMc regulation remains unknown at this time.

1.6: Dissertation Overview.

The major findings of this dissertation include (i) mapping the gene structure of mouse RGMc and genomic comparison to RGMc in other species, (ii) characterizing the promoter of RGMc in skeletal muscle, (iii) identifying a post-transcriptional regulatory element in the untranslated region of RGMc mRNA, and (iv) creating a phylogentic tree and *ab initio* protein model of the RGM family.

Chapter 2 will investigate the current understanding of the RGM family, including evolutionary relationships and *ab initio* modeling to query possible structure-function relationships of the RGM proteins, as well as identify areas for future research. In particular, the chapter highlights the fact that little is known about the regulatory

mechanisms of gene expression in the RGM family, especially for RGMc, and provides an introduction to the major gaps in knowledge that are addressed in this dissertation.

The aim of chapter 3 is to determine the regions of the RGMc locus that contribute to promoter activity during muscle differentiation. Data presented in this chapter demonstrate that (i) transcription of RGMc is induced early in muscle differentiation, and (ii) that three primary regions of the proximal promoter are important for RGMc promoter activity in differentiating muscle cells, being identified via a combination of promoter-reporter with site-directed mutagenesis experiments, gel-shift assays, and co-transfection/overexpression experiments to identify the most probable transcription factors that regulate these regions. Furthermore, a survey of the proximal promoter coupled with fragments of the RGMc locus that are well-conserved in mammals reveals all promoter activity is located within a region less than 1 kb.

Chapter 4 identifies a region of the RGMc mRNA 5' untranslated region (UTR) that increases luciferase reporter activity by 10-fold in muscle and over 40-fold in three different liver cells lines. *The primary focus* of the chapter will be to determine the fundamental mechanism by which this element operates and provide preliminary insight into future work on the regulation of RGMc expression. Three hypotheses are presented and experimental data currently support a mechanism of translational regulation.

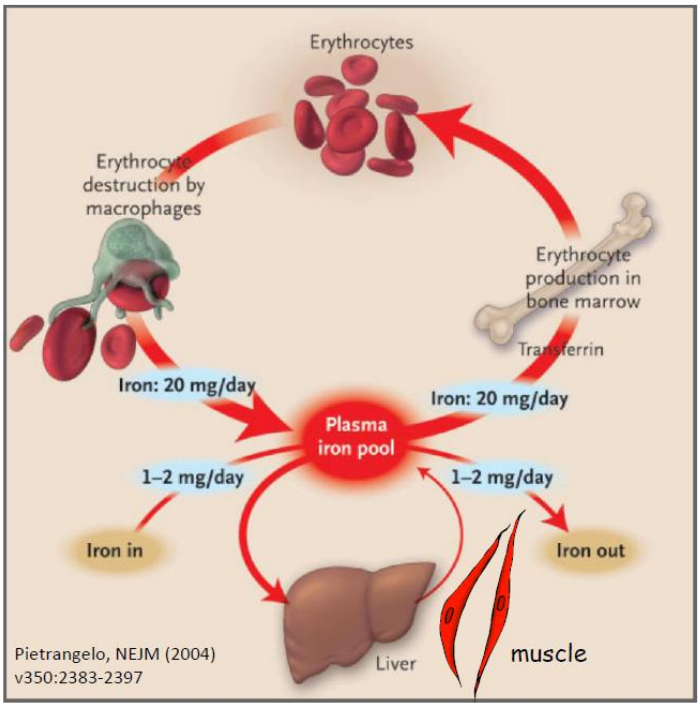
Finally, the appendix represents additional results related to the dissertation. Collectively, the data and discussion presented in this work create a detailed understanding of the most fundamental aspects of RGMc gene regulation and provide the foundation for future investigation into the evolution of gene families.

(This page was intentionally left blank)

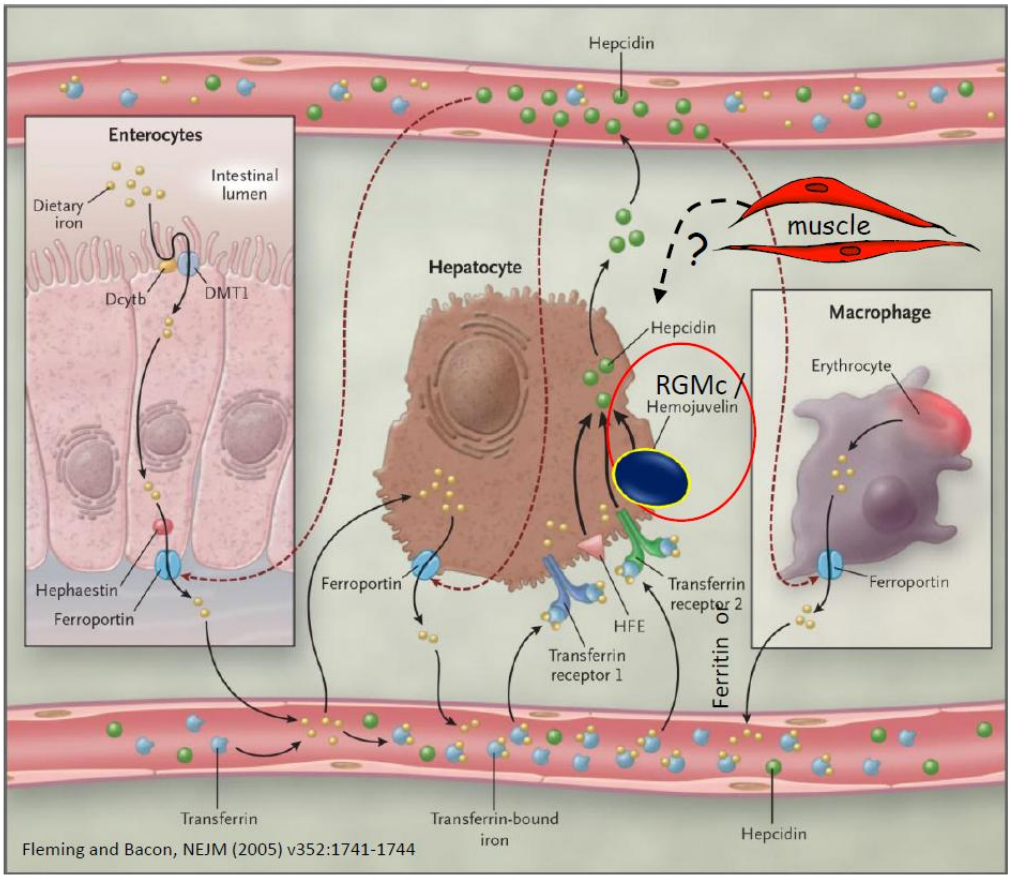
Figure 1.1: Models of the pathways involved in iron homeostasis.

Iron levels in the plasma are tightly regulated in multicellular organsisms. In humans, **A.** the majority of iron is found with erythrocytes (red blood cells) and hepatocytes (~1.0 g). Approximately 20 mg of iron is required daily in the bone marrow for incorporation into hemoglobin in erythroid precursor cells. Most of the iron found in the plasma derives from the continuous breakdown of hemoglobin in senescent red cells by reticuloendothelial macrophages. Approximately 1 to 2 mg per day is also taken up by duodenal enterocytes and transferred to the plasma compartment or, depending on body needs, stored in the enterocytes as ferritin. These stores are eliminated when enterocytes are sloughed off at the end of their life cycles; apart from menstrual blood loss, this is the only significant means by which excess body iron is excreted. Iron recycled by macrophages (as well as that absorbed from the gut) is loaded onto serum transferrin and delivered primarily to the bone marrow for reincorporation into new red-cell precursors. **B.** In the duodenal enterocyte, dietary iron is reduced to the ferrous state by duodenal ferric reductase (Dcytb), transported into the cell by divalent metal transporter 1 (DMT1), and released by way of ferroportin into the circulation. Hephaestin facilitates enterocyte iron release. Hepatocytes take up iron from the circulation either as transferrin-bound iron (through transferrin receptor 1 and transferrin receptor 2). Transferrin receptor 2 may serve as a sensor of circulating transferrin-bound iron, thereby influencing expression of the iron regulatory hormone hepcidin. The hepcidin response is also modulated by HFE and RGMc / Hemojuvelin, via an unknown mechanism, and to date, the source(s) of soluble RGMc is unknown. Hepcidin is secreted into the circulation, where it down-regulates the ferroportin-mediated release of iron from enterocytes, macrophages, and hepatocytes (dashed red lines). Figures and figure legends taken from Pietrangelo (2004), Ref. [24]; Fleming and Bacon (2005), Ref. [25].

A


Erythrocytes
Erythrocyte destruction by macrophages
Erythrocyte production in bone marrow
Transferrin
Iron: 20 mg/day
Iron: 20 mg/day
Plasma iron pool
1–2 mg/day
1–2 mg/day
Iron in
Iron out
Liver
muscle
Pietrangelo, NEJM (2004) v350:2383-2397

B


Hepcidin
Enterocytes
Intestinal lumen
Dietary iron
Dcytb
DMT1
Hephaestin
Ferroportin
Hepatocyte
muscle
?
Hepcidin
RGMc / Hemojuvelin
Macrophage
Erythrocyte
Ferroportin
Ferroportin
Transferrin receptor 2
HFE
Transferrin receptor 1
Ferritin or
Transferrin
Transferrin-bound iron
Hepcidin
Fleming and Bacon, NEJM (2005) v352:1741-1744

(This page was intentionally left blank)

Chapter 2

The Repulsive Guidance Molecule Family

"Nothing in biology makes sense except in the light of evolution." –Theodosius Dobzhansky

"Evolution does not produce novelties from scratch. It works on what already exists; either transforming a system to give it new functions or combining several systems to produce a more elaborate one." –François Jacob, Science (1977), Ref. [26]

The majority of the research in this chapter was originally published as:

Molecular biology, genetics and biochemistry of the repulsive guidance molecule family.

Christopher J. Severyn, Ujwal Shinde, and Peter Rotwein

Department of Biochemistry and Molecular Biology, Oregon Health & Science University, 3181 SW Sam Jackson Park Road, Portland, OR 97239-3098, U.S.A.

Biochemical Journal (2009) **422**, 393-403

Received 26 June 2009; accepted 6 July 2009

Published online 27 August 2009, doi:10.1042/BJ20090978

This work was supported by the National Institutes of Health, grant numbers R01 DK42748 (to P.R.), T32 HL007781 (Molecular Hematology Training Grant) and F30 HL095327 (to C. J.S.) and the National Science Foundation, grant number NSF-0746589 (to U.S.).

2.1 Summary

Repulsive guidance molecules (RGM) comprise a recently discovered family of glycosylphosphatidylinositol (GPI)-linked cell-membrane-associated proteins found in most vertebrate species. The three proteins, RGMa, RGMb, and RGMc, are products of distinct single-copy genes that arose early in vertebrate evolution, are ~ 40 - 50% identical to each other in primary amino acid sequence, and share similarities in predicted protein domains and overall structure, as inferred by *ab initio* molecular modeling, yet the respective proteins appear to undergo distinct biosynthetic and processing steps, whose regulation has not been characterized to date. Each RGM also displays a discrete tissue-specific pattern of gene and protein expression, and each is proposed to have unique biological functions ranging from axonal guidance during development (RGMa) to regulation of systemic iron metabolism (RGMc). All three RGM proteins appear capable of binding selected bone morphogenetic proteins (BMPs), and interactions with selected BMPs mediate at least some of the biological effects of RGMc on iron metabolism, but to date no role for BMPs has been defined in the actions of RGMa or RGMb. RGMa and RGMc have been shown to bind to the trans-membrane protein neogenin, which acts as a critical receptor to mediate the biological effects of RGMa on repulsive axonal guidance and on neuronal survival, but its role in the actions of RGMc remains to be elucidated. Similarly, the full spectrum of biological functions of the three RGMs has not been completely characterized yet, and will remain an active topic of ongoing investigation.

2.2: Introduction

The repulsive guidance molecule (RGM) gene family consists of three recently-discovered members, RGMa, RGMb, and RGMc [3, 27-31]. Each gene encodes a protein whose expression is restricted to a small number of tissues and is hypothesized to be involved in distinct biological functions ranging from control of iron metabolism to regulation of axonal guidance and neuronal survival in the developing nervous system. The RGM family receives its name from the axonal guidance molecule RGMa [28], a protein found primarily in the developing and adult central nervous system [27-29, 32]. A second member, RGMb (or Dragon [30]) is also detected in the nervous system, but in a different expression pattern than RGMa [30, 33]. The biological actions of RGMb are poorly characterized to date. The third member of the family is RGMc (also called hemojuvelin (HJV), HFE2, and Dragon-like muscle (DL-M)). Unlike RGMa or RGMb, RGMc is not expressed in the nervous system, but rather is produced by striated muscle and the liver [29, 31, 33, 34]. RGMc surprisingly regulates iron metabolism, as inactivating mutations cause juvenile hemochromatosis, a severe systemic iron overload disorder in humans [3]. To date, there has been no comprehensive assessment of the most fundamental aspects of the biology of the RGM family, including regulation of gene expression, control of protein biosynthesis, the relationship of protein structure to function, or mechanisms of action of each of the RGM proteins. In this chapter we address the molecular biology and biochemistry of the RGM family, attempt to define and critically evaluate what is known, and identify new areas for future investigation.

2.3: RGMa

2.3.1: Chromosomal organization and gene structure.

RGMa has been identified in ten mammalian and eight non-mammalian vertebrates, where it is a single-copy gene (Table 2.1). A single RGM gene also has been described in several invertebrate species, including urochordates, echinoderms, mollusks, and nematodes [35], as will be discussed in the molecular evolution section below. In vertebrates, RGMa comprises one of six conserved genes in a syntenic locus [36], as can be assessed by analysis of the corresponding parts of the human, mouse, and chicken genomes (Fig. 2.1). In these three species RGMa is positioned in the opposite transcriptional orientation from the other nearby genes. The locus is also conserved in zebrafish (Fig. 2.1). Within the cluster of six conserved genes near RGMc in human, mouse, and chick, Mctp2 (multiple C2 domains, transmembrane 2) is found 5' to RGMa, while Chd2 (chromodomain helicase DNA-binding protein 2), St8sia2 (ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 2), and Slco3a1 (solute carrier organic anion transporter family member 3A1) are located 3'. The latter three genes also are positioned downstream of RGMa in the zebrafish genome, but Mctp2 is absent (Fig. 2.1). In addition, in all four species, Nr2f2 (nuclear receptor subfamily 2, group F, member 2) is located upstream of RGMa, although both the relative orientation and the distance among species varies [~2 Mb in human and mouse genomes and ~830 kb in zebrafish, where the transcriptional direction is reversed] (Fig. 2.1).

Human and mouse RGMa genes are of comparable size, ~46 and ~44 kb respectively, and have a similar organization, being composed of four exons separated by three variably-sized introns, although the precise 5' end of exon 1 has not been defined in either species (Fig. 2.2 and Tables 2.2 and 2.3). In both genes, exon 1 is non-coding, and consists of most of the 5' un-translated region (UTR) of RGMa mRNA. Exon 2 contains the remaining 35 nucleotides of the 5' UTR and the first 26 codons of the RGMa protein, while exon 3 encodes the next 72 codons (73 in mice), and exon 4 the remaining 328 codons (321 in mice), plus a 3' UTR of ~1800 nucleotides and a single polyadenylation

signal (Fig. 2.2 and Table 2.3). The four exons are well conserved between human and mouse RGMa, with nucleotide identity ranging from a low of 64% for exon 1 to a high of 99% for exon 2 (calculated using refs. [37-40]). The three introns are less conserved than the exons (< 30% versus ~60% identity, respectively), although their lengths are similar between the two species (Fig. 2.2 and Table 2.3). Although four exons have been identified in the zebrafish RGMa gene, the nucleotide sequence of exon 1 is not similar to its mammalian counterparts [39-41]. In the chicken, the 5' end of the largest RGMa cDNA could not be mapped to the RGMa locus, possibly because the genomic sequence is incomplete in this region [42], and its DNA sequence also differs markedly from the other species. Thus, only three exons have been identified definitively in chicken RGMa, corresponding to mammalian exons 2 – 4 (Fig. 2.2).

2.3.2: Gene expression.

RGMa was cloned initially from mRNA isolated from chick embryonic optic tectum [28]. Subsequently, RGMa transcripts were shown to be expressed at highest levels in both the adult and developing central nervous system in chicken, mouse, and zebrafish [27-30, 32, 43]. RGMa mRNA has also been detected at lower levels in peripheral tissues, including heart, lung, liver, skin, kidney, and testis, at least in the adult rat [44]. By Northern blotting, the major RGMa transcript has been shown to be ~3.6 kb in length in the mouse [44], which is consistent with the aggregate size of the four RGMa exons [38, 45]. Other minor transcripts have been seen by Northern blotting, but their exact relationship with the RGMa gene has not been established to date [44, 46].

In the developing mouse embryo, RGMa mRNA has been detected as early as embryonic day (E) 8.5 in the neural folds of the central nervous system [27]. Later in development RGMa transcripts are found in several brain regions, including hippocampus, midbrain, the ventricular zone of the cortex, and parts of the brainstem and spinal cord [27, 33, 46].

Similar observations have been reported in the developing chicken [28, 32] and zebrafish [30]. The biochemical processes responsible for these distinct patterns of RGMa gene expression in the central nervous system have not been elucidated to date, in large part because nearly nothing is known about the organization or function of the RGMa gene promoter, about mechanisms of regulation of RGMa gene transcription, or about RGMa mRNA turnover. Similarly, the signaling pathways that govern RGMa gene expression in different tissues and in response to physiological and pathological stimuli have not been characterized.

2.3.3: Protein sequence and expression.

The initial identification of chick RGMa after its cDNA cloning revealed it to be a cell membrane-associated glycosylphosphatidylinositol (GPI)-linked two-chain protein that was derived from a primary translation product of 432 amino acids [28]. Subsequent cloning of human and mouse RGMa cDNAs predicted similarly sized proteins of 434 and 438 residues [27], respectively, that were 91% identical to each other and 80% identical to chick RGMa (Table 2.4). In all three species and in zebrafish RGMa, the NH_2-terminal signal peptide is estimated to be ~30 residues, although the first amino acid of the mature protein has not been characterized experimentally. The RGMa precursor also contains a conserved GPI attachment signal at its COOH-terminus of ~45 amino acids. This segment is removed in the endoplasmic reticulum during RGMa biosynthesis when the GPI anchor is added to the nascent protein [28, 47]. Other recognizable protein elements in RGMa include an RGD motif (a potential integrin binding site [28, 48], for sequence details, see Fig 2.15), and a partial von Willebrand type-D domain (vWD) [28, 49] (Fig. 2.16 for sequence conservation) that contains the site of internal cleavage to generate two-chain RGMa (Fig. 2.3) [these domains and other aspects of the biochemistry of RGM proteins will be discussed in the section on structure – function relationships]. The mechanism of intramolecular cleavage of RGMa has not been

established, although it appears to occur during its biosynthesis, leading to a mature RGMa that is a disulfide-bonded two-chain protein composed of an NH_2-terminal fragment of ~123 residues, and a COOH-terminal segment of ~238 residues [28, 50], and that is linked to the outer face of the plasma membrane by its COOH-terminal GPI anchor [28, 51, 52] (Fig. 2.3B). The number and pattern of disulfide bonds has not yet been established for the 14 cysteines found in mature RGMa [a molecular model is discussed in the section on structure – function relationships]. RGMa also appears to be a glycoprotein, with three potential asparagine-linked glycosylation sites in mammals and two in the chicken (Fig. 2.3A and Fig. 2.4) [28, 51]. At present it is not known if other RGMa isoforms exist, such as single-chain species, or whether soluble forms of the protein are found in the extra-cellular fluid.

2.3.4: Physiological functions and mechanisms of action.

RGMa was identified as a factor involved in guiding axons by repulsion from the temporal half of the developing chicken retina toward the anterior optic tectum in the brain, and membranes derived from cells expressing chick RGMa were shown to inhibit temporal retinal growth cones, but had little effect on nasal growth cones [28]. Perhaps surprisingly given these initial observations, genetic knockout of RGMa in mice did not alter retinal axonal patterning, but rather caused defects in neural tube closure [27]. Thus, the exact *in vivo* functions of RGMa in mammals remain to be determined.

It has been shown that RGMa regulates repulsive guidance of retinal axons via binding to neogenin [32, 53] (summarized in Fig. 2.5), a trans-membrane protein that is also a receptor for netrins, a family of secreted molecules involved in neuronal development and cell survival (reviewed in [54]). Unlike netrins, RGMa does not bind to proteins related to neogenin, such as DCC (deleted in colorectal cancer), or members of the Unc (uncoordinated) sub-family [53], although recent observations suggest an indirect

association with Unc5b [55, 56]. In addition to regulating retinal axonal guidance, the interaction between RGMa and neogenin has been found to promote neuronal survival [32]. Initial studies of the early events triggered after RGMa binds to neogenin have suggested the involvement of several signal transduction intermediates, including protein kinase C, the small GTPase RhoA, RhoA kinase [52, 57], and focal adhesion kinase [55, 56], as well as the putative transcriptional co-activator, LIM-only protein 4 [58], but the full spectrum of biochemical mechanisms responsible for mediating the biological effects of RGMa by neogenin has not been established.

Like other members of the RGM family, RGMa has been found to bind to selected bone morphogenetic proteins (BMPs) [44, 59], which belong to the TGF-β growth factor family [60]. In initial biochemical studies, a fusion protein composed of human RGMa linked to the immunoglobulin G (IgG) Fc fragment was shown to bind radiolabeled BMP-2 and BMP-4, but not BMP-7 or TGF-β1 in cross-linking experiments [59]. In cell-based studies, over-expression of RGMa was found to increase activity of a co-transfected promoter-reporter gene containing a BMP response element (BRE), while knockdown of endogenous RGMa led to a reduction in reporter gene expression [59]. Although these preliminary observations are intriguing, a role for BMPs in the biological actions of RGMa has not been defined.

2.4: RGMb

2.4.1: Chromosomal organization and gene structure.

RGMb is a single-copy gene in the eight mammalian and seven non-mammalian vertebrates in which it has been identified (Table 2.1). Like RGMa, RGMb resides within a conserved chromosomal locus, and comprises one of five linked genes that are found in the same relative orientation to each other in the human, mouse, and chicken

genomes (Fig. 2.6). In each of these species, RGMb is located in a tail-to-tail transcriptional orientation with Chd1 (chromodomain helicase DNA-binding protein 1), in a relationship similar to that of RGMa and Chd2 (compare Figs. 2.6 and 2.1). This suggests that a duplication event involving this chromosomal region occurred during evolution prior to the emergence of mammals. Further away and upstream of RGMb are Riok2 (right open reading frame kinase 2), Lix1 (Limb expression 1) and Lnpep (leucyl/cystinyl aminopeptidase) (Fig. 2.6). In contrast, very little is currently known about the chromosomal environment of RGMb in the zebrafish genome (Fig. 2.6).

The human RGMb gene is ~25 kb in length, and contains 5 exons (Fig. 2.7, Tables 2.2 and 2.6), including two 5' non-coding exons (1 and 2), which include ~406 nucleotides of a ~524 nucleotide 5' UTR of RGMb mRNA. The 5' end of exon 1 has not been mapped. The remaining 118 nucleotides of the 5' UTR are found in exon 3, which also includes the first 45 codons of the coding region. Exon 4 encodes the next 170 codons, and exon 5 the remaining 222 codons plus a 3' UTR of 308 nucleotides that includes a single polyadenylation signal (Fig. 2.7 and Table 2.6). In the mouse genome, only three RGMb exons have been identified to date, and these correspond to exons 3 - 5 of the human RGMb gene (Fig. 2.7). The 3' UTR of mouse RGMb mRNA encoded by exon 3 is longer than its human counterpart, being ~2.5 kb in length. In zebrafish, only the coding region for RGMb has been mapped to its genome [30], and is found within three distinct exons (Fig. 2.7).

2.4.2: Gene expression.

RGMb was discovered by an informatics-based search for genes related to RGMa [27], and was independently cloned as a gene whose putative promoter was bound by the homeodomain transcription factor, DRG11, which is expressed in dorsal root ganglia (DRG) of the sympathetic nervous system [30, 61, 62]. RGMb (DRG-'ON' or Dragon)

was co-localized with DRG11 mRNA in dorsal root ganglia and in the spinal cord. RGMb mRNA was also detected in the developing neural tube prior to the onset of expression of DRG11, and has been found in other areas of the nervous system where DRG11 is not produced [30]. This latter result suggests that RGMb gene expression is controlled by additional regulatory factors besides DRG11. Results of *in situ* hybridization experiments show that RGMb mRNA is expressed in the DRG, in the spinal cord excluding the ventricular zone, in the retina, in the optic nerve, and in other distinct regions of the brain, including the developing mouse midbrain, hindbrain, and forebrain [27, 30, 33, 63], although the pattern of RGMb gene expression does not overlap appreciably with that of RGMa [27]. RGMb mRNA also has been detected in the nervous system of the developing zebrafish [30], and has been found in the reproductive tract of rodents [64]. Based on results of Northern blotting studies, there appears to be a single RGMb transcript in mice of ~4.2 kb [27, 30], which is approximately the same size as the three mouse RGMb exons (Table 2.2). As with RGMa, the mechanisms responsible for RGMb gene expression in different tissues or under different physiological or pathological conditions have not been characterized, and virtually nothing is known about the structure or function of the RGMb gene promoter.

2.4.3: Protein sequence and expression.

Cloning of mouse RGMb cDNA revealed a predicted protein of 438 amino acids [27, 30], which is 89% identical to human RGMb (437 residues long) and 65% identical to zebrafish RGMb (436 amino acids) (Table 2.4). The primary RGMb translation product is predicted to contain an NH_2-terminal signal peptide of ~50 residues, although this has not been verified experimentally, and a COOH-terminal GPI attachment signal of ~35 amino acids [27, 30]. Other identifiable motifs in RGMb include a partial vWD element. After forced expression of mouse RGMb in Hek-293 and COS-7 cells, only a single protein band of ~50 kDa could be detected in cell extracts by immunoblotting, and a

similarly sized protein was released into the culture medium after incubation of cells with phosphatidyl-inositol phospholipase C (PI-PLC), which cleaves the GPI anchor [27, 30]. These latter results indicate that only a single-chain RGMb species is attached to the outer face of the cell membrane [30, 65] (Fig. 2.3B), although the protein contains a putative internal proteolytic cleavage site similar to that in RGMa. RGMb also appears to be glycoprotein and is predicted to encode up to two asparagine-linked glycosylation sites (Figs. 2.3A and 2.4). As with RGMa, mature RGMb contains 14 cysteines whose potential organization into disulfide bonded residues has not been established [but see discussion of potential molecular models in the section on structure –function relationships].

2.4.4: Potential physiological functions.

No biological functions of RGMb have been elucidated except for its possible ability to promote cell – cell adhesion by homophilic interactions [27, 30], and its capability to bind selected BMPs [65, 66] (Fig. 2.5). As with RGMa, over-expressed full-length RGMb has been found to increase the activity of a promoter - reporter gene containing a BMP-responsive transcriptional control element in cell culture systems [64, 65], but unlike RGMa, RGMb has not been shown to bind to neogenin.

2.5: RGMc/Hemojuvelin

2.5.1: Chromosomal organization and gene structure.

RGMc is a single-copy gene in the nine mammalian and six non-mammalian vertebrates in which it has been identified (Table 2.1). Unlike RGMa and RGMb, RGMc has not been found to date in the chicken or other avian species. In human and mouse genomes, RGMc comprises one of 10 linked genes in a syntenic locus that includes among others, Txnip (thioredoxin interacting protein), Polr3gl (polymerase (RNA) III (DNA directed)

polypeptide G like), Ankrd34 (ankyrin repeat domain 34), Lix1l (related to Lix1, which maps near RGMb), and Chd1l (related to Chd1 and Chd2, which are located near RGMb and RGMa, respectively [compare Figs. 2.1, 2.6, and 2.8]). Of note, however, the relative transcriptional orientation of RGMc and Chd1l (tail-to-head) differs from that of RGMa - Chd2 and RGMb - Chd1 (tail-to-tail). Moreover, in zebrafish the RGMc chromosomal environment differs from mammals (Fig. 2.8). Although the location of two Txnip-like genes and Polr3gl are adjacent to RGMc (similar to what is seen in mammals), Mtx1 and Thbs3a are just upstream of zebrafish RGMc, while in mouse they are located at a distance of more than 8 Mb from RGMc. Furthermore, there is no Chd-homolog present on the zebrafish RGMc locus. Interestingly, zebrafish chromosomes 16 and 19 have a large duplicated region sharing parologous genes that flank RGMc on either side (Fig. 2.8), but an RGMc paralog is not found on chromosome 19.

Human and mouse RGMc genes are similar in size (~4.3 and ~4.0 kb, respectively, Table 2.2) and organization, being composed of four exons separated by three introns (Fig. 2.9), and are considerably smaller than mammalian RGMa or RGMb (Table 2.2). In both species, exon 1 is ~160 nucleotides in length [mapping experiments defining the 5' end of RGMc in mouse skeletal muscle, heart, and liver may be found in Chapter 3 of this Dissertation], and contains most of the 5' UTR of RGMc mRNA. The remaining 90 nucleotides of the 5' UTR are found in exon 2, along with the first 31 codons of the RGMc protein (28 in mouse). Exon 3 encodes the next 173 codons (169 in mouse), and exon 4 the remaining 222 codons (223 in mouse), plus a 3' UTR of ~1150 nucleotides with a single polyadenylation signal (Fig. 2.9 and Table 2.7). The four RGMc exons are well-conserved between the mouse and human genes, with nucleotide sequence identity ranging from 73 to 83% (calculated using refs. [37-40], see Table 2.7). The three introns are less conserved, although their lengths are similar between mouse and human (Fig. 2.9). The zebrafish RGMc gene is larger than its mammalian counterparts, and contains

5 exons distributed over ~ 11.4 kb (Fig. 2.9). Exons 1 and 2 are non-coding but are not similar in DNA sequence to mammalian RGMc exon 1. In contrast, zebrafish exons 3 – 5 correspond to mammalian RGMc exons 2 – 4, with nucleotide sequence identity ranging from 50 to 59% (coding and non-coding regions separated in Table 2.7).

2.5.2: Gene expression.

RGMc was independently discovered as a gene within a locus linked to the human iron overload disorder, juvenile hemochromatosis [3], as an mRNA related to RGMa and RGMb [27, 29, 30, 33], and as a novel transcript expressed during skeletal muscle differentiation [31]. In addition to skeletal muscle, RGMc mRNA has been detected in the heart and in the liver [27, 31, 33]. During mouse development, RGMc transcripts are found first in the somites, precursors of skeletal muscle, as early as E11.5, which is before muscle can be identified morphologically [31]. Similar observations have been made in zebrafish [30, 41]. In the mouse, RGMc mRNA is detected by E13.5 in the heart and liver [1, 31].

Very little is known about RGMc gene regulation. Elucidating the regulatory mechanisms of RGMc expression is the primary focus of this Dissertation. Chapters 3 and 4 will present the first evidence for regulation of RGMc at the level of transcription and provide supporting evidence for control at the level of translation. In mice, RGMc mRNA levels were shown to be decreased in the liver, but not in skeletal or cardiac muscle after systemic injection of bacterial lipopolysaccharide (Fig. 4 of Ref. [1]), but as with RGMa and RGMb, the biochemical mechanisms responsible for controlling RGMc gene transcription or mRNA stability in different tissues or under different physiological or pathological conditions had not been fully established when work for this thesis began. Chapter 3 will present experimental evidence for the structure and function of the RGMc

gene promoter and chapter 4 will examine the post-transcriptional regulatory mechanisms of RGMc expression.

2.5.3: Protein sequence, processing, and expression.

The initial cloning of human and mouse RGMc cDNAs revealed primary translation products of 426 and 420 amino acids, respectively, with a predicted NH_2-terminal signal peptide of ~31 residues and a COOH-terminal GPI-attachment signal of ~45 amino acids [27, 29, 34], although as in other RGM molecules, the precise boundaries have not been determined experimentally. Mouse and human RGMc precursor proteins are 88% identical to each other (Table 2.4). Like RGMa, RGMc contains up to three asparagine-linked glycosylation sites (Fig. 2.4), and like its paralogues, has several shared protein motifs, including an RGD sequence and a partial vWD domain with a conserved proteolytic cleavage site (Fig. 2.3A). In addition, and unlike RGMa or RGMb, mammalian RGMc proteins encode a furin-like pro-protein convertase (PPC) recognition and cleavage sequence near the COOH-terminus (Figs. 2.3 and 2.5), and the protein has been shown to be cleaved by furin at this site [67-69]. As a consequence, RGMc appears to undergo a complex series of biosynthetic and processing steps, leading to the production of four distinct protein isoforms (Fig. 2.5) in skeletal muscle and after expression of the recombinant protein in heterologous mammalian cells [34, 67, 69, 70]. Two of the RGMc isoforms (a disulfide-bonded two-chain species that is similar to RGMa, and a single-chain isoform similar to RGMb), are attached to the extra-cellular face of the plasma membrane by a GPI linkage [34, 67, 69, 71] (Fig. 2.3B). In addition, single-chain RGMc species have been detected in the extra-cellular fluid of cultured cells, and in blood [34, 67-72] (Fig. 2.3B). These latter two proteins differ at their COOH-termini, with the smaller species being derived from the larger by PPC-mediated proteolytic cleavage [34, 67, 69] (Fig. 2.5). Results of biosynthesis experiments further support the idea that the two soluble single-chain RGMc proteins originate from the

single-chain cell-associated molecule [34, 67] (Fig. 2.5). Analogous studies have not been reported for RGMa or RGMb. As in RGMa and RGMb, the disulfide bonding pattern of the 14 cysteines found in mature full-length RGMc has not been experimentally defined, but a possible model is discussed below (see Figs. 2.13 and 2.14).

2.5.4: Physiological functions and mechanisms of action.

A role for RGMc in systemic iron metabolism was first inferred when mutations in the human gene were linked to the severe iron overload disorder, juvenile hemochromatosis [3]. This relationship was strengthened when mice engineered to lack RGMc were found to have excessive accumulation of iron in multiple tissues [1, 73]. It has been postulated that the normal biological action of RGMc is to induce expression of the secreted hepatic peptide, hepcidin [1, 3], which is a negative regulator of dietary iron uptake from the duodenum and release of stored iron from macrophages [3, 74] (Fig. 1.1B). Humans with juvenile hemochromatosis and mice with RGMc deficiency have low levels of serum or urinary hepcidin [75, 76], and mice lacking RGMc also have diminished expression of hepcidin mRNA in the liver [1, 73]. The mechanism of regulation of hepcidin by RGMc is currently under active investigation with the leading hypothesis being that cell-membrane associated RGMc facilitates signaling by BMPs through its receptors to promote hepcidin gene expression [66, 77-79]. In this model, soluble RGMc has been proposed to act as an inhibitor, presumably by sequestering BMPs away from cell-surface receptors [69, 72].

Like RGMa, RGMc binds to the extra-cellular part of neogenin [70, 71, 80], although the role of neogenin in the biological actions of RGMc has not been established. One report has demonstrated preferential binding of two-chain RGMc to neogenin [70] (Fig. 2.5), while murine versions of two juvenile hemochromatosis-associated RGMc amino acid

substitution mutants, D172E and G320V (Fig. 2.10 and Table 2.8), which did not form a two-chain species [34, 70], were unable to bind [70]. Similar results were observed with the human G320V juvenile hemochromatosis-associated protein [34, 67, 69, 71]. In other experiments, neogenin was unable to alter BMP-mediated hepcidin gene expression [79], although it is unclear which RGMc protein isoforms were used in these studies. More work is needed to elucidate the biochemical mechanisms by which RGMc regulates systemic iron metabolism under different physiological conditions, to determine if there is a role for neogenin in the biological actions of RGMc, and to characterize the functions of different RGMc species in normal physiology and in disease.

2.6: Molecular Evolution of the RGM Family

One unresolved question about the RGM family concerns the evolutionary relationships among the three members. To address this issue we performed a series of phylogenetic analyses by querying multiple sequence alignments of selected RGM proteins after applying the following two criteria: (i) using only well-annotated sequences in which the protein defined by translation from both mRNA and genomic sequences is identical, and (ii) minimizing the level of 'mammalian bias' by selecting RGM genes from a diversity of organisms. We found that three-quarters of our assessments supported the hypothesis that RGMc diverged from a common ancestor earlier than did RGMa or RGMb (see legend to Fig. 2.12 for a summary of methods). Two of the phylogenetic trees are presented in Fig. 2.12. Similar conclusions were reached by Schmidtmer and Engelkamp [29], while Camus and Lambert have advocated the alternative viewpoint that RGMa and RGMc are more closely related to one another [35].

Inspection of RGM genomic loci strengthens the view that RGMa and RGMb have a closer relationship to each other than to RGMc. RGMa and RGMb genes are physically

linked to Chd2 and Chd1, respectively, in mammalian, chicken, and zebrafish genomes (Figs. 2.1 and 2.6), and are each part of a more extensive syntenic linkage group that includes in order (at least in the human genome) RGMA - CHD2 - ST8SIA1 - SLCO3a1 and RGMB - CHD1 - ST8SIA4 - SLCO4C1, indicating that the organization of paralogous genes within the duplicated chromosomal regions has been maintained (Figs. 2.1 and 2.6). By contrast, only a Chd1-related pseudo-gene is found near the same chromosomal locus as RGMc in mammals, but is located at a much greater distance from RGMc than Chd2 or Chd1 are from RGMa or RGMb, respectively (compare Figs. 2.1, 2.6, 2.8). Also in mammals the pseudo-gene, Lix1-like, is found near RGMc, but in a different relationship than Lix1 and RGMb (compare Figs. 2.6 and 2.8). Interestingly, all known RGMc genes use three exons for the protein coding sequence (CDS) (Fig. 2.11), despite a variable number of overall exons (which alter the untranslated regions (UTR)) (Fig. 2.9). As shown in figure 2.11, RGMc uses one nucleotide of the codon from the first CDS exon and two nucleotides in the second exon to complete the codon triplet (Fig. 2.11). The junction between the second and third CDS exons are separate codons (*green* to *red* in Fig. 2.11). This pattern is well-conserved from mammals (humans and mice) to bony fish (zebrafish) and also reflected in the nucleotide conservation between mammals and fish of the middle CDS exon (see 'exon 3' in Table 2.7). Major differences in sequence conservation occurs in the CDS that codes for the N-linked glycosylation "γ-site" (Fig. 2.4), which is unique to mammals. It should be noted that there are no known JH-causing mutations within this putative N-linked glycosylation site (Fig. 2.10 and Table 2.8). The final CDS exon contains the region that encodes the PPC cleavage site (Fig. 2.3), which is only found in mammalian RGMc and not RGMc in bony fish and amphibians, and both nucleotide (Table 2.7) and protein conservation reflect this difference.

Single RGM genes have been identified in several invertebrates. The evidence is strongest for existence of an RGM protein in the sea squirt, *Ciona intestinalis*, where a polyadenylated mRNA has been characterized that corresponds to the four-exon genomic DNA sequence (NCBI accession number AK173741), and encodes a predicted protein of 637 amino acids (calculated using Transeq [40]), with multiple cysteine residues (15 in the putative mature protein vs. 14 in vertebrate RGMs), and overall similarity of 40%, 38%, or 27% to mouse RGMa, RGMb, or RGMc, respectively. Like RGMb, *Ciona* RGM contains no RGD motif, but instead has a RGN sequence [40, 81] (Fig. 2.15). Similar to mammalian RGMc, the *Ciona* RGM has a predicted PPC site near its COOH-terminus. As *Ciona* is a model organism used extensively in development, it will be interesting to see where this gene is expressed. For example, if the single RGM is expressed in muscle-structures of the developing and mature *Ciona*, this would greatly strengthen the hypothesis that expression of RGMs in muscle is an evolutionary ancient phenomena (discussed in detail in chapter 3). To date however, this putative gene and protein has not been characterized.

An RGM gene also has been identified in the purple sea urchin, *Strongylocentrotus purpuratus*, where it maps near a CHD1-like gene (LOC575959) as seen in RGMa and RGMb loci in vertebrates (Figs. 2.1 and 2.6). The protein predicted to be encoded by this gene contains an RGD motif (Fig. 2.15) and 16 cysteines (14 of which align with the 14 conserved cysteines in mammalian RGMs), and is ~40% identical to mammalian RGMa or RGMb, and ~35% identical to RGMc [82]. In the nematode, *Caenorhabditis elegans*, a single RGM gene also has been predicted, but the putative protein is < 30% identical to mammalian RGMs, lacks several of the conserved cysteine residues found in mammalian RGM proteins, and unlike vertebrate RGM proteins does not contain either an RGD or RGN sequence [83]. Although a single RGM has been reported in mollusks (California brown sea slug, *Aplysia californica*) [35], definitive genomic evidence is lacking.

Clearly, further analysis of putative RGM genes and their encoded proteins in invertebrates is needed for more complete understanding of the evolution and functions of the RGM family.

2.7: Structure-Function Relationships among RGM proteins

Three dimensional structures can provide critical insights to structure-function relationships within a protein family. Although no such information is available yet for the RGM family, emerging computational methods such as comparative modeling [84, 85], fold recognition[86], and *ab initio* techniques [87, 88] have the potential to help overcome this deficiency. Comparative modeling can approximate the three-dimensional structure of a target protein for which only the amino acid sequence is available, provided that an empirical three-dimensional "template" structure is available from a protein with >30% sequence identity. Alternatively, threading methods, which search for an optimal fit of query sequences onto known three-dimensional structures of proteins in databases, can be used when a comparative modeling approach is unsuccessful. However, neither comparative modeling nor threading techniques were able to identify appropriate templates for RGM proteins. As a consequence, we constructed initial structural models for the RGM family with *ab initio* approaches, which use the physical properties of the primary amino acid sequence to predict structures. We employed 'Rosetta' *ab initio* modeling software because it has been the most consistent and accurate in predicting structures of folded domains in a series of trials (CASP: critical assessment of techniques for protein structure predictions [87-94]). For the RGM family, structural segments were generated using the Rosetta fragment server with input amino acid sequence information derived from 22 RGM proteins (see legend to Fig. 2.14). One thousand independent simulations were generated and were organized into clusters according to structural

similarities, as outlined in the legend to Fig. 2.14. All *ab initio* models analyzed suggest that RGM proteins adopt a two-lobed structure (Fig. 2.14).

Mature RGMa, RGMb, and RGMc each contain 14 similarly placed cysteine residues (Fig. 2.3A), and all appear to be disulfide-bonded proteins [34, 69, 71]. However, the number and location of disulfide bonds are unknown. The majority of *ab initio* models show a disulfide bond between cysteine-9 and either cysteine-7 or -8 (Fig. 2.13), although one model suggests two disulfide bonds (Fig. 2.14A, cysteines shown as space-filling models in purple), and this could be the linkage responsible for maintaining two-chain forms of RGMa or RGMc. Both cysteine-11 and -12, and cysteine-13 and -14 also are predicted to form disulfide bonds in all models generated, and are located within the COOH-terminal part of the two-lobed structure (Fig. 2.14A). While the connectivity varies slightly between models (Fig. 2.13), the majority of the predictions suggest two disulfide bonds for the NH_2-terminal lobe between cysteines-1 and -2, and cysteines-4 and -5, for a total of 5 or 6 disulfide linkages per RGM molecule. This would leave 2 - 4 free cysteines in the protein (Fig. 2.14A). Clearly, direct experiments are needed to define the actual disulfide bonding pattern for each RGM family member.

Von Willebrand factor is a glycoprotein that helps mediate platelet adhesion in hemostasis (i.e., arrest of bleeding) at damaged blood vessels through interactions with blood clotting factor VIII [49, 95]. It contains five distinct structural domains (vWA, B, C, D, CK) [49], and one of these motifs (type D) has been recognized in all RGM proteins [29]. The vWD-domain is a structured region in which the first and second conserved cysteines form a disulfide bridge [49, 95], and this native conformation of the vWD-domain is required for factor VIII binding, as well as normal multimerization of the vWF protein [95]. Our *ab initio* models suggest that this partial vWD domain is highly structured, and contains surface exposed α-helices and β-strands (yellow region in Fig.

2.14). These are consistent with the crystal structure of the entire vWD domain (RCSB protein structural data base accession #1ijb) [96]. Whether this structure is required for multimerization of RGMc is not known at this time, but it is intruging to speculate about the possible function of this well-conserved domain. The RGM partial vWD region contains the site of intramolecular proteolytic cleavage to generate two-chain forms of RGMa and RGMc (see Figs. 2.3 and 2.5), and this cleavage has been hypothesized to occur by acid-labile hydrolysis between an aspartic acid and proline residue [71]. In the model depicted in Fig. 2.14, these two amino acids are located on the surface of the protein (surface of space-filling model in 2.14B), between a highly conserved DP (Asp-Pro) doublet found in all RGM family members (Fig. 2.16). By comparison, in the mammalian protein zonadhesin (ZAN), which contains five tandem vWD-domain repeats, three of the vWD repeats contain an epidermal growth factor (EGF)-like motif [**] [97] followed by a "DP" doublet that are cut [98]. In the absence of the EGF-like motif and "DP" doublet, no cleavage occurs in ZAN [98]. Interestingly, there is another conserved "DP-site" in RGMb and RGMc, located to the N-terminal side of the vWD that does not appear to be used for cleavage, as well as a third "DP" found in mammalian RGMc adjacent to the N-linked glycosylation site (sequence –NFT–), though these sites do not contain the consensus sequence GDPH suggested to be used in the mucin protein family for auto-catalysis at the "DP-site" [99]. In contrast, the vWD-domain of RGMb does not appear to be cleaved, suggesting that this proteolytic site used to generate the two-chain form of RGMa and RGMc requires a specific recognition through an enzymatic cleavage event rather than strictly undergoing auto-proteolysis under acidic conditions. The only predicted enzyme to cleave at such a site is the *Staphylococcus aureus* Protease-V8 (also called Endoproteinase GluC), leaving the mammalian enzyme

[**] A full epidermal growth factor (EGF)-like domain contains six conserved cysteines involved in three disulfide bonds, as well as a two-stranded beta-sheet followed by a loop to a C-terminal short two-stranded sheet. EGF-like motifs have variable number of cysteines. Cell (1996) v**85**(4):597-605

capable of cleaving this site unknown to date. Alternatively, subtle differences in accessibility or local pH concentrations of the "DP-site" between RGM family members may support the auto-proteolysis model for the cleavage seen in RGMa/c, but not in RGMb.

The only appreciable differences between RGMb and its paralogs within the linear sequence of this region are (i) a stretch of ~15 amino acids unique to RGMb homologs just before the partial vWD, (ii) a phenylalanine (Phe, F) in RGMb instead of a histidine (His, H) residue before the predicted site of cleavage in RGMa and –c which could cause the cleavage site to be buried and therefore inaccessible, (iii) an acidic glutamic acid (Glu, E) in the place of a glutamine (Gln, Q) found in RGMa and –c, and (iv) two sites where RGMb contains a Ser instead of Gln (RGMa) or hydrophobic residue (RGMc), or where RGMa and RGMc contain a Ser/Thr, which is a Val in RGMb (Fig. 2.15). Whether these differences in the linear vWD sequence alter the predicted cleavage site is not known. Site-directed mutagenesis studies could readily differentiate between these possibilities. Furthermore, a difference in the three-dimensional structure could also be a reason for the difference between the lack of internal cleavage in the RGMb form when compared to its paralogs, and thus the critical residues may be quite distant on the linear protein sequence. For example, the G320V mutation in RGMc that results in a single chain form of RGMc that is not cleaved, and is unable to bind neogenin like the two-chain form [70], is ~150 amino acids away from the cleavage site in the linear sequence. Understanding the three dimensional structure of the family will likely provide critical insights into the nature of processing and interacting partners of the RGM family.

Of note, a substitution of this aspartic acid residue to glutamic acid in human RGMc (D172E) causes juvenile hemochromatosis [100], and in biochemical experiments the mutant protein does not form a two-chain molecule [34, 70]. Another disease-causing amino acid substitution in human RGMc of glycine 320 to valine (G320V) also appears

to block production of the two-chain protein [34, 70]. The *ab initio* model depicted in Fig. 2.14 suggests that G320 is located on a surface that is in proximity to D172. On the basis of the model it thus appears possible that the G320V substitution, which increases the side-chain volume and hydrophobicity, may inhibit interactions with some unknown protein/protease to prevent proteolysis at residue D172. Alternatively, the substitution may induce certain conformational changes that indirectly impair proteolytic cleavage at D172.

RGMa and RGMc each contain a RGD motif (arginine-glycine-aspartic acid), a tripeptide classically identified as an integrin-binding element [48], while RGMb does not [29, 48] (Fig. 2.15). Structurally, RGD motifs are found at or near the end of an α-helix [101], and our *ab initio* models map the RGM RGD sequence to a loop between two α-helices on the surface of the protein (Fig. 2.14A). The exact function of this motif in RGMa or RGMc is not known, although amino acid substitutions of glycine to valine or arginine (G99V or G99R) appear to cause juvenile hemochromatosis in humans [3, 100], and the analogously mutated mouse RGMc (G92V) was unable to bind BMP-2 in biochemical assays [70].

RGM proteins contain several putative asparagine-linked glycosylation sites (Fig. 2.4), and have been shown to be glycoproteins [28, 34, 51], although the functional role of glycosylation has not been established for any RGM family member yet. In our *ab initio* structural models, at least two of these sites map to the surface of the molecule (Fig. 2.14). There are however, two conserved $–NX^{S}/_{T}–$sequence motifs in the N- and C-termini of all RGM family members (except bony fish), a unique motif found before the partial vWD only in RGMa (–NYT–), and a unique motif after the partial vWD in mammalian RGMc (–NAT–). The only known disease-causing mutation in a glycosylation site can be found in RGMc, where a conserved Cys is mutated to a Phe (–

NCS–), however this is likely to be the result of structural modifications to the Cys residue rather than ablation of a glycosylation site, as the consensus motif is still retained as a putative glycosylation site. Interestingly, the "DP" motifs in RGMb and RGMc that do not appear to be cleaved are adjacent to $–NX^{S}/_{T}–$ sequence motifs and potential mucin-type O-linked glycosylation sites. Nevertheless, O-linked glycans are not added at a defined consensus motif and their presence cannot be accurately predicted based on the primary sequence alone [102]. Furthermore, it has not been shown that any RGM family members contain O-linked glycosylation sites to date, but recent advances in O-linked-glycan-techniques may illuminate this question [102]. Future work to determine whether all of these putative sites are glycosylated in the RGM family members may shed light on their relevance to the structure and function of the proteins. As noted earlier, RGMc but not RGMa or RGMb contains a pro-protein convertase recognition and cleavage site near the COOH terminus of the mature protein (Fig. 2.3). As seen in Fig. 2.14A, this part of the protein in our *ab initio* model also maps to a surface loop, and thus potentially would be readily accessible to targeted proteolysis by furin or other pro-protein convertases.

2.8: Summary and Challenges for the Future

The RGM family appears to have been composed of three genes early in vertebrate evolution, being present in a common ancestor to mammals and fish. Each gene is expressed in a distinct developmental and tissue-specific pattern, with RGMa and RGMb being produced in different parts of the central nervous system, and RGMc being synthesized in striated muscle and liver. The molecular mechanisms governing such diverse tissue-restricted gene expression have not been established, and little is known about the structure or function of RGM gene promoters, about their mechanisms of transcriptional regulation, or about control of RGM mRNA processing or stability.

Details of theses processes of gene regulation for one of the family members, RGMc/Hemojuvelin, will be presented in chapters 3 and 4, contributing to our understanding of RGM gene expression. At the protein level, the three RGM family members share several motifs and are predicted to have similar three-dimensional structures based on our *ab initio* modeling, but the respective proteins appear to undergo distinct biosynthetic and processing steps, whose regulation has not been characterized. From the perspective of function, all three RGM proteins appear capable of binding selected BMPs, although binding domains have not been mapped. It appears that interactions with selected BMPs may mediate at least some of the biological effects of RGMc to control hepcidin gene expression, but to date no role for BMPs has been defined in the actions of RGMa or RGMb. To date only RGMa and RGMc have been shown to bind to neogenin, and while signaling through neogenin is critical for the biological effects of RGMa on repulsive axonal guidance and on neuronal survival, its role in the actions of RGMc remains to be elucidated. Similarly, the full spectrum of biological functions of the three RGMs has not been completely characterized yet, and will remain an active topic of ongoing investigation.

2. 9: Acknowledgements

We thank Kevin Kendall at MacVector, as well as David Kuninger and Lisa Wilson for advice and guidance. In addition, preliminary aspects of Table 2.8 were developed by David Kuninger.

No conflict of interest to report.

Table 2.1: Species in which more than one RGM has been identified [a]

Species (abbreviation)		**RGMa**	**RGMb**	**RGMc**
Mammals				
Human	*Homo sapiens (Hsa)*	AK074910 AL136826	BC067736	AK223575 AK092682
Chimpanzee	*Pan troglodytes (Ptr)*	+	+	+
Rhesus macaque	*Macaca mulatta (Mmul)*	+	+	+
Pig	*Sus scrofa (Sscr)*	+	--	--
Dog	*Canis familiaris (Cfa)*	+	--	+
Cow	*Bos taurus (Bta)*	+	+	+
Elephant	*Loxodonta africana (Laf)*	+	+	--
Mouse	*Mus musculus (Mmu)*	BC059072 BC023870	AK047390 BC096024	AJ557515
Rat	*Rattus norvegicus (Rno)*	+	--	BC089203
Armadillo	*Dasypus novemcinctus (Dno)*	--	+	+
Opossum	*Monodelphis domestica (Mdo)*	+	+	+
Non-mammalian vertebrates				
Chicken	*Gallus gallus (Gga)*	AY128507	+	--
Frog	*Xenopus tropicalis (Xtrop)*	BC061329	BC061325	+
Zebrafish	*Danio rerio (Dre)*	BC091800 AY613931[b]	AY613929	BC134888 BC112964
Salmon	*Salmo salar (Ssa)*	BT045779	--	--
Japanese Pufferfish	*Takifugu rubripes (Tru)*	+	+	+
Green-spotted Puffer	*Tetraodon nigroviridis (Tni)*	+	+	+
Stickleback	*Gasterosteus aculeatus (Gac)*	+	+	+
Medaka (Killer fish)	*Oryzias latipes (Ola)*	+	+	+

a. Accession numbers for cDNAs listed. All others have been identified through homology mapping in their respective genomes.
b. Mis-labeled in GenBank as DL-M (muscle RGMc).

Table 2.2: Characteristics of RGM genes

	Species	Gene Size (kb)	# Exons	mRNA (kb)
RGMa	Human	45.8	4	3.2
	Mouse	44.4	4	3.6
	Zebrafish	12.5	4	4.5
RGMb	Human	24.8	5	2.2
	Mouse	20.3	3	4.2
	Zebrafish	18.3	3	> 1.3
RGMc	Human	4.3	4	2.1
	Mouse	4.0	4	2.0
	Zebrafish	11.4	4	1.7

Table 2.3: RGMa gene characteristics

	Gene Size (kb)	mRNA (kb)	Exon 1 [a] nt	Exon 1 [a] (% identity)	Intron 1 kb	Exon 2 nt	Exon 2 (% identity)	Exon 2 nt	Exon 2 (% identity)	Intron 2 kb	Exon 3 nt	Exon 3 (% identity)	Intron 3 kb	Exon 4 nt	Exon 4 (% identity)	Exon 4 nt	Exon 4 (% identity)
Human	45.8	3.2	286	(64)	15.9	34	(100)	82	(99)	20.4	515	(90)	6.3	705	(83)	1594	(61)
Mouse	44.4	3.6	369	(100)	15.5	35	(100)	82	(100)	17.7	518	(100)	7.6	714	(100)	1867	(100)
Zebrafish	12.5	4.5	478[b]	(<40)	2.1	163	(<40)	82	(68)	4.1	518	(66)	1.8	699	(64)	2580	(40)
Chicken	unk	>1.5	unk	--	--	49	(<40)	79	(76)	unk	515	(80)	1.8	702	(74)	141	(<30)

a Exons numbered and % identity in relation to mouse RGMa. Non-coding regions in *yellow*; coding sequences in *blue*; unk = unknown

b Zebrafish exon 1 is implied from EST data (accession AL911518 and EH589480)

Table 2.4: Amino acid identity among RGM proteins

			% *vs.* Mouse		
	Species	Size (aa)	RGMa	RGMb	RGMc
RGMa	Human	434	91	50	48
	Mouse	438	100	49	48
	Zebrafish	433	68	48	43
	Chicken	432	80	53	45
RGMb	Human	437	52	89	42
	Mouse	438	49	100	42
	Zebrafish	436	46	65	42
RGMc	Human	426	49	44	88
	Mouse	420	48	42	100
	Zebrafish	410	44	41	46

Calculations are based on Smith-Waterman (local) alignment using Blosum62 matrix, gap open penalty of 10.0, and gap extend penalty of 0.5. The protein sequences are derived from cDNAs whose accession numbers are listed in Table 2.1.

Table 2.5: Abbreviations in Genomic Loci

Abbreviation	Definition
NR2F2	nuclear receptor subfamily 2, group F, member 2
MCTP	multiple C2 domains, transmembrane
RGM	Repulsive guidance molecule
CHD	chromodomain helicase DNA binding protein 2
ST8SIA2	ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase
SLCO3A1	solute carrier organic anion transporter family, member 3A1
GNRH-R4HS	gonadotropin-releasing hormone receptor GnRH-R4SHS
LNPEP	leucyl/cystinyl aminopeptidase
LIX1	protein limb expression 1
RIOK2	right open reading frame kinase 2
PRDM9	PR domain containing 9
TXNIP	thioredoxin interacting protein
POLR3GL	polymerase (RNA) III (DNA directed) polypeptide G like
ANKRD34	ankyrin repeat domain 34
MTX1	metaxin 1
THBS3	thrombospondin 3
RBM8A	RNA binding motif protein 8A
PEX11B	peroxisomal biogenesis factor 11 beta
ZNF364	Zfp364 (zinc finger protein 364)
FCGR1A	Fc fragment of IgG, high affinity Ia, receptor (CD64)

Table 2.6: RGMb gene characteristics

	Gene Size	mRNA	*Hsa* Exons [a]		Exon (1) [b]				Intron (1)	Exon (2)		Intron (2)	Exon (3)			
	(kb)	(kb)	nt	(% identity)	nt	(% identity)	nt	(% identity)	kb	nt	(% identity)	kb	nt	(% identity)	nt	(% identity)
Human	24.8	2.2	331+75 [a]	(<30)	118	(89)	136	(83)	5.4	509	(88)	13	663	(87)	313	(85)
Mouse	20.3	4.2	unk	--	377	(100)	145	(100)	4.8	509	(100)	13.5	660	(100)	2474	(100)
Zebrafish	18.4	>1.3 [d]	unk	--	unk	--	106	(44)	12.7	524	(70)	4.3	678	(66)	unk	--

a Human (*Hsa*) exons 1 and 2 listed together as 331+75 (exon 1 is 331 bp, exon 2 is 75 bp, with a 0.97 kb intron separating exons 1-2, and 3.3 kb intron separating exons 2-3). % identity for *Hsa* exons in Human exons 1 and 2 compared to the mouse 5' UTR, separately and together. Overall identity is less than 30%.

b Exons tentatively numbered to reflect that the exact number of exons for RGMb is not known. Exons numbered and % identity in relation to mouse RGMb. Non-coding regions in *yellow*; coding sequences in *blue*; unk = unknown

c 42 nt at the 5'end of the cDNA do not align with the human genomic sequence

d Coding sequence only for zebrafish

Table 2.7: RGMc gene characteristics

	Gene Size (kb)	mRNA (kb)	Exon 1 [a] nt	Exon 1 [a] (% identity)	Intron 1 kb	Exon 2 nt	Exon 2 (% identity)	Exon 2 nt	Exon 2 (% identity)	Intron 2 kb	Exon 3 nt	Exon 3 (% identity)	Intron 3 kb	Exon 4 nt	Exon 4 (% identity)	Exon 4 nt	Exon 4 (% identity)
Human	4.3	2.1	156	(75)	1.3	90	(77)	97	(84)	0.40	560	(83)	0.47	621	(85)	612	(59)
Mouse	4.0	2.0	160	(100)	1.1	86	(100)	88	(100)	0.43	548	(100)	0.48	624	(100)	524	(100)
Zebrafish	11.4	1.7	65+58[b]	(46)	1.4	74	(55)	103	(45)	1.4	512	(59)	4.90	615	(50)	280	(40)

a Exons numbered and % identity in relation to mouse RGMc. Non-coding region in *yellow*; coding sequence in *blue*; unk = unknown

b Zebrafish exons 1 and 2 listed together as 65+58 (exon 1 is 65 bp, exon 2 is 58 bp, with a 2.0 kb intron separating the exons)

Table 2.8 Disease-causing mutations in human Hemojuvelin (HJV; human RGMc)

WT	Residue	Mutation	Genetics	Potential Pathophysiology	RGM Conservation	Reference(s)
Gln	6	His	a	Intracellular mis-sorting?		*103*
Cys	80	Arg	a	Unpaired Cys? (or altered disulfide bond) Model suggests unpaired Cys	**	*104*
Ser	85	Pro	b	Structural Change (alpha-helix disruption)	*	*100*
Gly	99	Val	b	RGD interruption	**	*3*
Gly	99	Arg	a	RGD interruption	**	*100*
Leu	101	Pro	a, b	RGD interruption; breaking a-helix	**	*104*
Cys	119	Phe	b	Altered disulfide bond? Alpha-helix disruption @ -NCS- site	**	*105*
Ala	168	Asp	b	Altered internal cleavage? (DP at 172-173)	**	*100*
Phe	170	Ser	b	Altered internal cleavage? (DP at 172-173)	**	*100*
Asp	172	Glu	a	Altered internal cleavage? (DP at 172-173)	**	*100*
Arg	176	Cys	a	Altered internal cleavage? (DP at 172-173) Altered disulfide bond?	**	*106*
Trp	191	Cys	b	internal cleavage site (beta-sheet disruption)	**	*100*
Ser	205	Arg	a	RGMb) Predicted to be adjacent to DP site	**	*100*
Ile	222	Asn	a	Structural Change (alpha-helix disruption) Structural Change in vWD	**	*3*
Asp	249	His	b	Structural Change (alpha-helix disruption)	**	*107*
Gly	250	Val	a	Altered flexibility in region predicted to be between N-term and C-term	**	*100*
Ile	281	Thr	a, b	Structural Change (beta-sheet disruption)	**	*3, 103*
Arg	288	Trp	b	Structural Change (beta-sheet disruption)	**	*100*
Gly	320	Val	a, b	Altered structure around Cys 321?	**	*3, 100, 105-109*
Cys	321	Trp	a	Loss of disulfide bond?	**	*110*
Leu	27	fsX	b	Not synthesized?	*	*111*
Arg	54	fsX	b	Not synthesized?		*112*
Gly	66	fsX	b	Not synthesized?		*113*
Val	74	fsX	a	Not synthesized?		*100*
Gln	116	fsX	a	Not synthesized?	**	*109*
Arg	131	fsX	b	Not synthesized?		*100*
Asp	149	fsX	b	Not synthesized?		*100*

(continued on next page)

Table 2.8 Disease-causing mutations in human Hemojuvelin (HJV; human RGMc) (continued from previous page)

WT	Residue	Mutation	Genetics	Potential Pathophysiology	RGM Conservation	Reference(s)
Asn	269	fsX	a	Not synthesized?		*100*
Gln	312	fsX	b	Misfolded protein?	**	*107*
Gly	319	fsX	a	Misfolded protein?	*	*100*
Cys	321	fsX	a	Unpaired Cys? (Altered disulfide bond) No GPI anchor	**	103
Arg	326	fsX	a	No GPI anchor PPC disruption?		3, 108
Ser	328	fsX	a	No GPI anchor	*	*105*
Cys	361	fsX	b	Unpaired Cys? (Altered disulfide bond) No GPI anchor	**	*3*
Arg	385	fsX	b	No GPI anchor	*	*100*

Genetics: a = Compund Heterozygote b = Homozygous

Relative RGM Conservation: ** = Highly Conserved * = low conservation

fsX = frameshift

References: [3, 100, 103-113]

Figure 2.1: Comparative structures of RGMa genomic loci. The relative position of the RGMa gene (red line) is indicated on each chromosome (Chr.; human 15, mouse 7, chicken 10, zebrafish 18) in relation to the centromere (grey oval, if information available) and telomere. Presented below each chromosome is a higher resolution view of the RGMa locus for each species. Neighboring genes are indicated, with the transcriptional direction represented by an arrow. Gene names corresponding to the abbreviations may be found in Table 2.5.

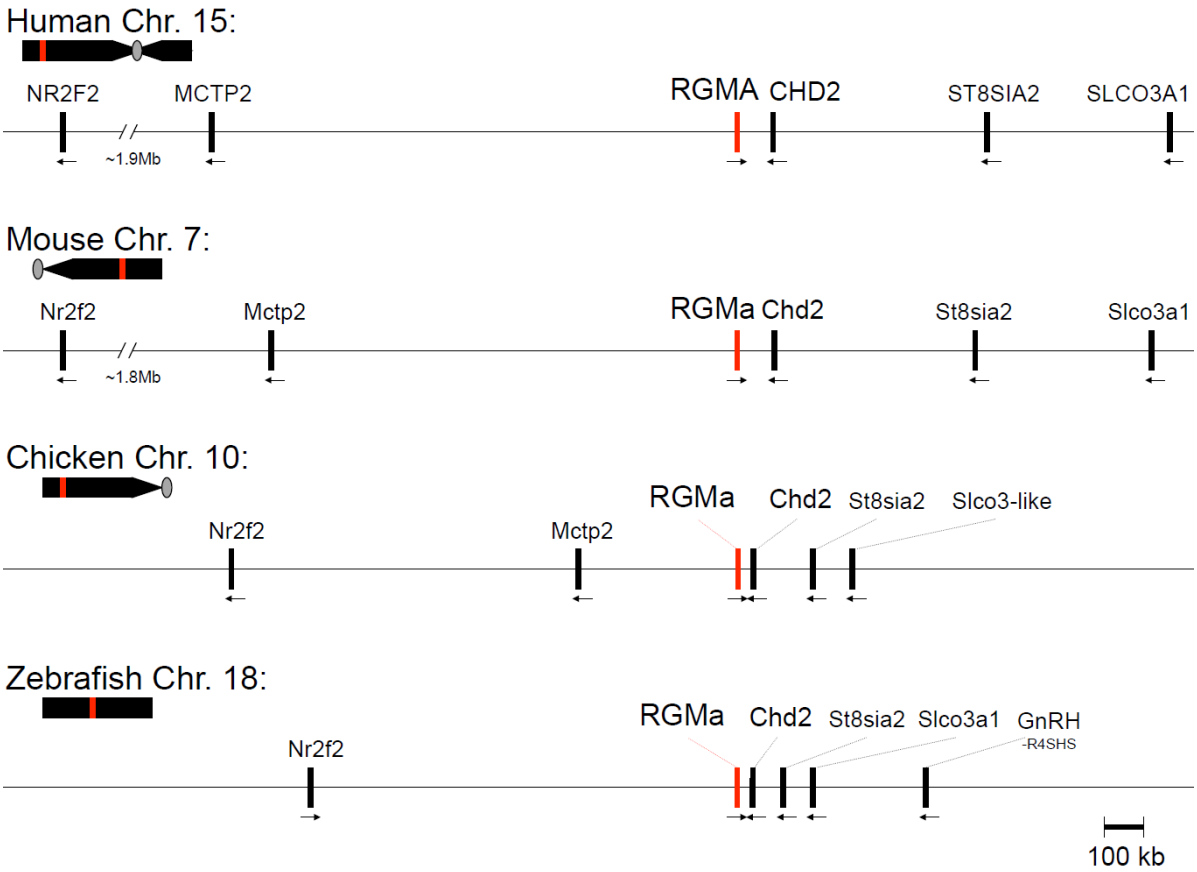

Human Chr. 15:
NR2F2
MCTP2
~1.9Mb
RGMA
CHD2
ST8SIA2
SLCO3A1
Mouse Chr. 7:
Nr2f2
Mctp2
~1.8Mb
RGMa
Chd2
St8sia2
Slco3a1
Chicken Chr. 10:
Nr2f2
Mctp2
RGMa
Chd2
St8sia2
Slco3-like
Zebrafish Chr. 18:
Nr2f2
RGMa
Chd2
St8sia2
Slco3a1
GnRH
-R4SHS
100 kb

Figure 2.2: Comparative organization of RGMa genes. The anatomy of human, mouse, zebrafish, and chicken RGMa genes is shown. Exons are indicated by *boxes*, with coding regions in *blue* and non-coding regions in *yellow*. The assignment of exon numbers is based on comparison with mouse RGMa. The polyadenylation site when known is depicted by a *vertical arrow*. The location of zebrafish exon 1 is based on mapping available EST data taken from GenBank (accession numbers AL911518 and EH589480). The length of one of the introns of chicken RGMa is not known (shown as *two angled lines*), as putative exon 2 cannot be mapped to the genomic DNA sequence, which appears to be incomplete in this region. Chicken exon assignments are in parentheses because putative exon 1 cannot be mapped to the genome.

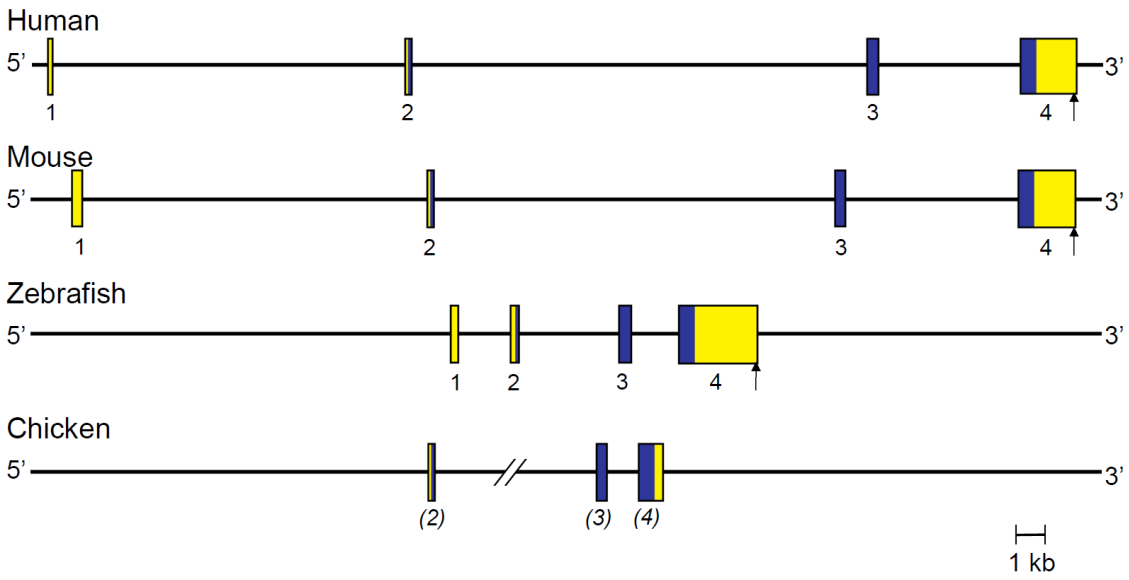

Human
5'
3'
1
2
3
4
Mouse
5'
3'
1
2
3
4
Zebrafish
5'
3'
1
2
3
4
Chicken
5'
3'
(2)
(3)
(4)
1 kb

Figure 2.3: Characteristics of RGM proteins. **A**. The linear maps of mature RGMa, RGMb, and RGMc contain the following features: RGD (arginine-glycine-aspartic acid) motif (RGMa and RGMc - green); vWD, partial vonWillebrand type-D domain (yellow); PPC, pro-protein convertase recognition and cleavage site (RGMc only - purple); * - location of asparagine-linked glycosylation sites; vertical solid arrowhead - site of intra-molecular proteolytic cleavage to generate two-chain RGMa and RGMc; vertical open arrowhead - possible site of intra-molecular proteolytic cleavage in RGMb; solid red vertical lines - conserved cysteine residues. The squiggle at the COOH-terminus of each protein represents the GPI anchor. **B**. Schematic of mature RGMa, RGMb, and RGMc on the cell surface, as well as the secreted forms of RGMc. Based on published studies, RGMa appears to be primarily a two-chain molecule, and RGMb a single-chain protein, while RGMc appears to be represented by both single- and two-chain species. Experimental data supports at least one disulfide bond between the NH_2- and COOH-termini [34, 69, 71], and *ab inito* molecular modeling (see Fig. 2.14) predicts one or two disulfide bonds connecting the two-chain RGM isoforms (shown as –S-S–), though the exact number is currently unknown. Single chain RGMc is released from the cell surface, and is found in extracellular fluid and in blood [34, 67-72], potentially through the actions of a furin-like pro-protein convertase (PPC) and/or a phosphatidylinositol phospholipase C (PI-PLC). It is not known if RGMa, RGMb, or two-chain RGMc are released from the membrane (as indicated by arrows with question marks). Locations of asparagine-linked glycosylation sites are indicated by asterisks, and the GPI-anchor is depicted as a squiggle.

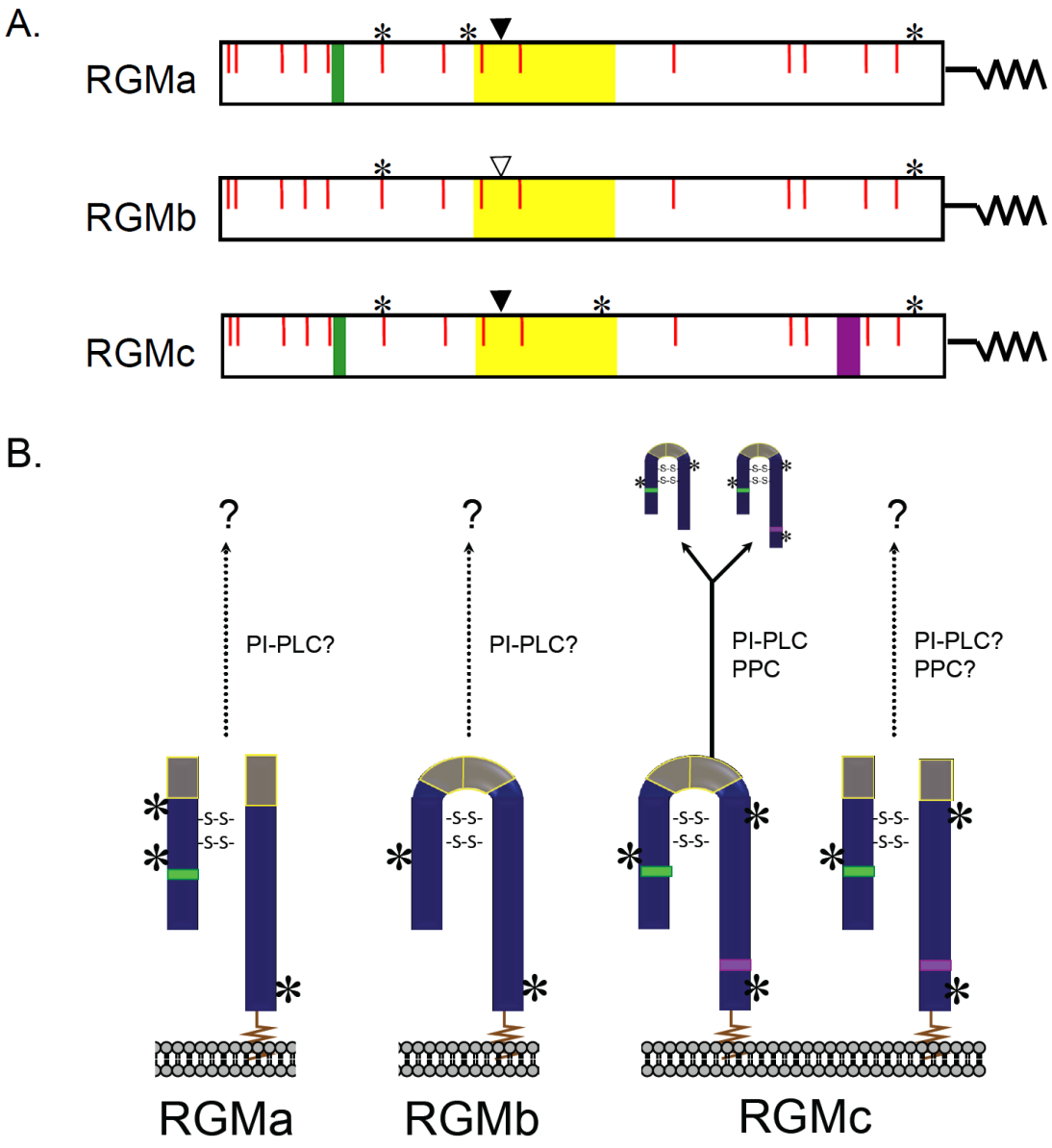
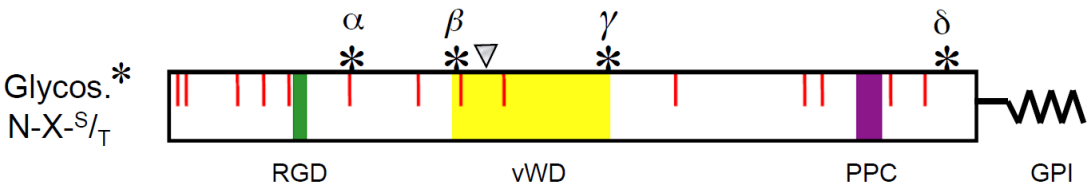

A.
RGMa
RGMb
RGMc
B.
?
?
?
PI-PLC?
PI-PLC?
PI-PLC
PPC
PI-PLC?
PPC?
-S-S-
-S-S-
RGMa
RGMb
RGMc

Figure 2.4: Comparative structures of putative N-linked glycosylation sites in the RGM family. The linear maps of mature protein for the RGM family highlighting the asparagine (Asn; N)-linked glycosylation sites, depicted as an asterisk, *. In addition, the following features are noted: RGD (arginine-glycine-aspartic acid) motif (RGMa and RGMc - green); vWD, partial vonWillebrand type-D domain (yellow); PPC, pro-protein convertase recognition and cleavage site (RGMc only - purple); vertical arrowhead - site of intra-molecular proteolytic cleavage to generate two-chain RGMa and RGMc, and a possible site of intra-molecular proteolytic cleavage in RGMb; solid red vertical lines - conserved cysteine residues. The squiggle at the COOH-terminus of each protein represents the GPI anchor. '+' indicates the motif is present (but may or may not be experimentally determined to be an N-linked glycosylation site) in all species noted in Table 2.1, unless otherwise noted, '-' indicates motif not present in those species surveyed.



	α-site -NCS-	β-site -NYT-	γ-site -NAT-	δ-site -NXS/T-
RGMa	+	+	-	+ Not in chicken
RGMb	+	-	-	+
RGMc	+ Not in bony fish	-	+ Not in bony fish	+

Figure 2.5: Model of the generation of different known RGM family isoforms. A. Schematic of the generation of the four possible isoforms known to be present in at least one of the RGM family members. Identifiable motifs are as follows: RGD (arginine-glycine-aspartic acid) motif (RGMa and RGMc - green); vWD, partial vonWillebrand type-D domain (yellow); PPC, pro-protein convertase recognition and cleavage site (RGMc only - purple); squiggle at the COOH-terminus of each protein represents the GPI anchor; S-S, are disulfide bonds, the exact number of which are unknown to date; scissors indicate a likely enzymatic cleavage point, based on predictions in [71, 98]. **B**. The different known isoforms based on experimental data and the known interactions between BMP-2 or neogenin, with relative strength of binding indicated as a '+' for stronger interactions, or '-' with no appreciable interactions shown based on data from Refs. [70]; '?' indicates a hypothesized interaction based on studies of similar isoforms from other RGM family members. Asparagine (Asn; N)-linked glycosylation sites are omitted from **A**. for clarity, and included as an asterisk, '*', in **B**. Figure is modeled after Kuninger, et al., *BMC Biochemistry* (2008) v**9**(1):9, and Kuns-Hashimoto, et al., *Am J Physiol Cell Physiol* (2008) v294(4):C994-C1003.

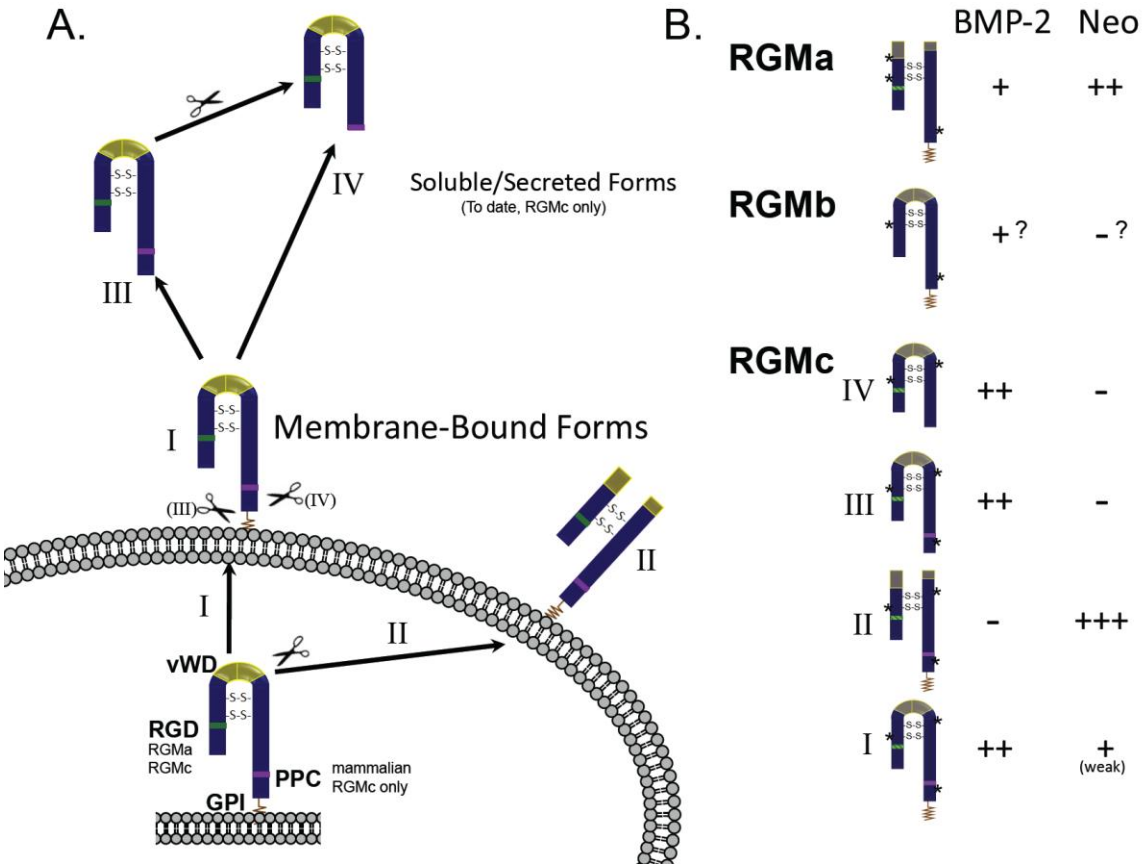

A.
IV
Soluble/Secreted Forms
(To date, RGMc only)
III
I
Membrane-Bound Forms
(III)
(IV)
II
I
II
vWD
RGD
RGMa
RGMc
PPC
mammalian
RGMc only
GPI
B.
BMP-2
Neo
RGMa
+
++
RGMb
+ ?
- ?
RGMc
IV
++
-
III
++
-
II
-
+++
I
++
+
(weak)

Figure 2.6: Comparative structures of RGMb genomic loci. The relative position of the RGMb gene (red line) is indicated on each chromosome (Chr.; human 5, mouse 17, chicken Z, zebrafish 5) in relation to the centromere (grey oval, if information available) and telomere. Presented below each chromosome is a higher resolution view of the RGMb locus for each species. Neighboring genes are indicated, with their transcriptional direction represented by an arrow. For the zebrafish RGMb locus, a nearby provisional gene is shown in grey; to date no other genes have been mapped to this region. Gene names corresponding to the abbreviations may be found in Table 2.5.

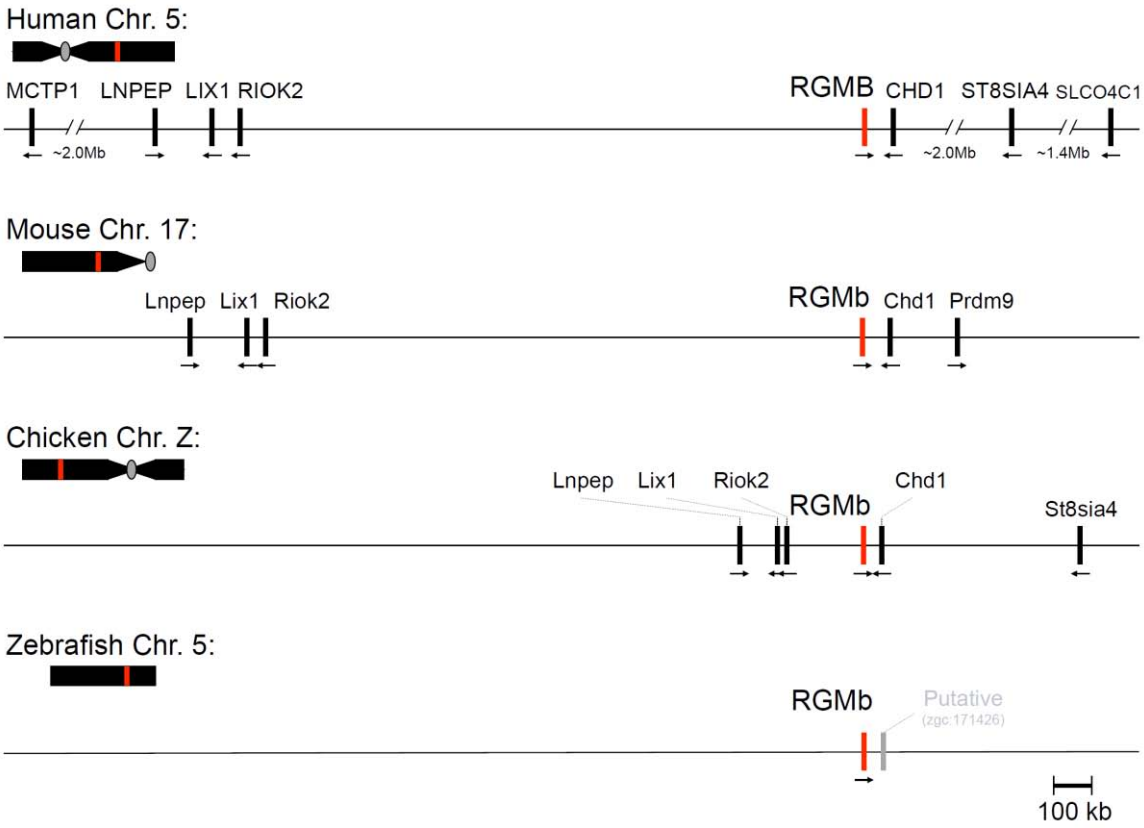

Human Chr. 5:
MCTP1
LNPEP
LIX1
RIOK2
~2.0Mb
RGMB
CHD1
ST8SIA4
SLCO4C1
~2.0Mb
~1.4Mb
Mouse Chr. 17:
Lnpep
Lix1
Riok2
RGMb
Chd1
Prdm9
Chicken Chr. Z:
Lnpep
Lix1
Riok2
Chd1
RGMb
St8sia4
Zebrafish Chr. 5:
RGMb
Putative
(zgc:171426)
100 kb

Figure 2.7: Comparative organization of RGMb genes. The anatomy of human, mouse, and zebrafish RGMb genes is shown. The assignment of exon numbers is based on comparison with human RGMb, and is provisional for mouse and zebrafish, as indicated by the parentheses. Exons are indicated by *boxes*, with coding regions in *blue* and non-coding regions in *yellow*. The polyadenylation site, when known, is depicted by a *vertical arrow*. Only coding information is available for zebrafish RGMb.

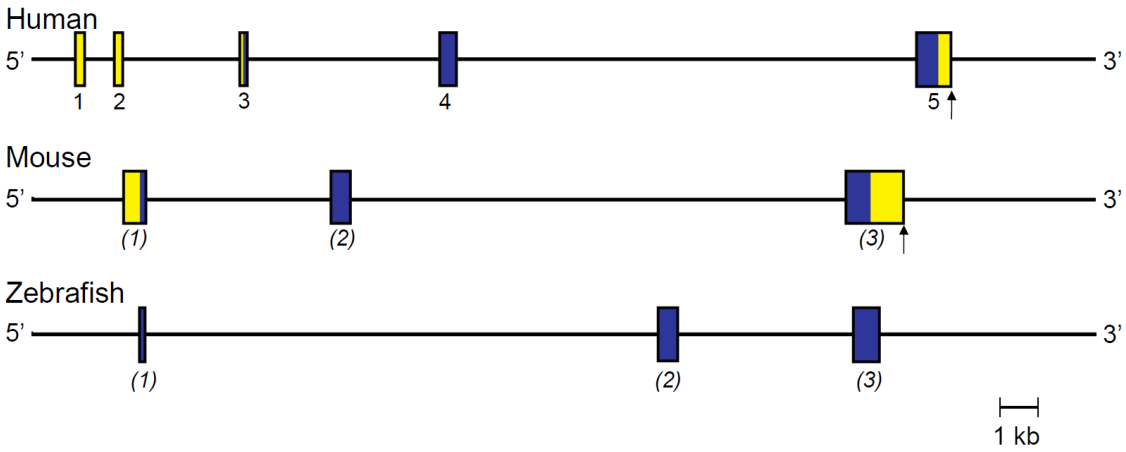

Human
5'
3'
1
2
3
4
5
Mouse
5'
3'
(1)
(2)
(3)
Zebrafish
5'
3'
(1)
(2)
(3)
1 kb

Figure 2.8: Comparative structures of RGMc genomic loci. The relative position of the RGMc gene (red line) is indicated on each chromosome (Chr.; human 1, mouse 3, zebrafish 16 and 19) in relation to the centromere (grey oval, if information available) and telomere. Presented below each chromosome is a higher resolution view of the RGMc locus for each species. Neighboring genes are indicated, with their transcriptional direction represented by an arrow. Lix1-like and Polr3gl-like are putative-pseudo-genes (Lix1l and Polr3gl-l), as there is no known transcript available in GenBank. A region of zebrafish Chr. 19 is shown to indicate a duplication of genes when compared to Chr. 16, but RGMc has not been duplicated on Chr. 19. Gene names corresponding to the abbreviations may be found in Table 2.5.

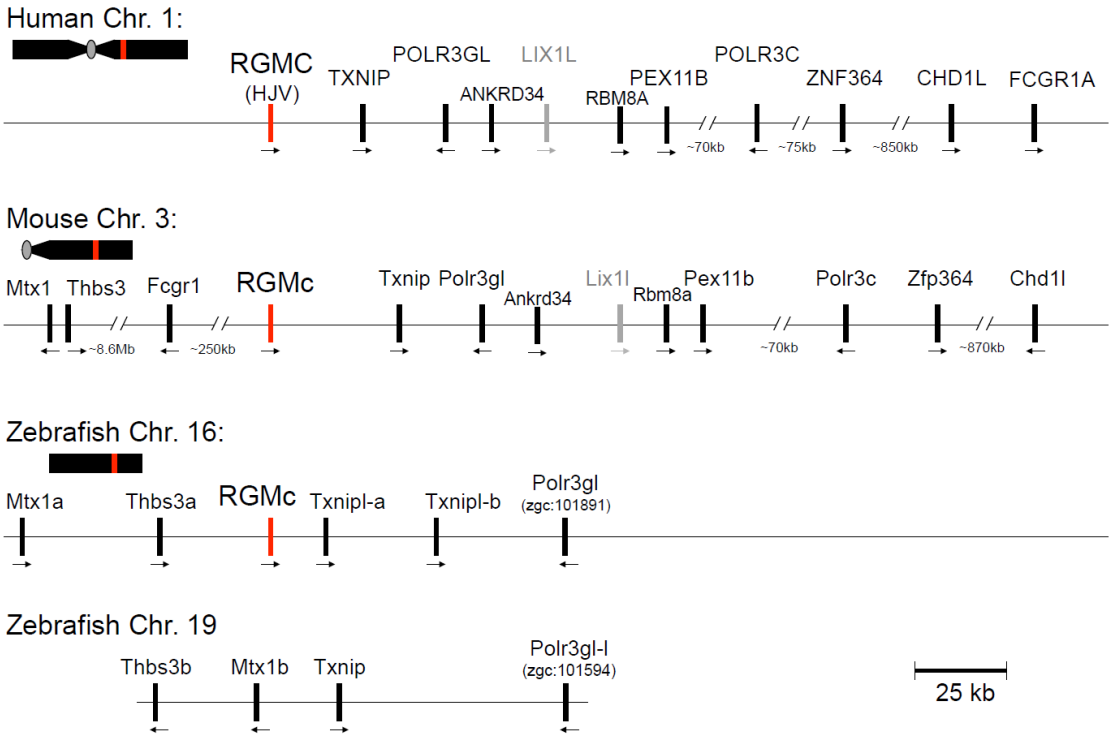

Human Chr. 1:
RGMC
(HJV)
TXNIP
POLR3GL
ANKRD34
LIX1L
RBM8A
PEX11B
POLR3C
ZNF364
CHD1L
FCGR1A
~70kb
~75kb
~850kb
Mouse Chr. 3:
Mtx1
Thbs3
Fcgr1
RGMc
Txnip
Polr3gl
Ankrd34
Lix1l
Rbm8a
Pex11b
Polr3c
Zfp364
Chd1l
~8.6Mb
~250kb
~70kb
~870kb
Zebrafish Chr. 16:
Mtx1a
Thbs3a
RGMc
Txnipl-a
Txnipl-b
Polr3gl
(zgc:101891)
Zebrafish Chr. 19
Thbs3b
Mtx1b
Txnip
Polr3gl-l
(zgc:101594)
25 kb

Figure 2.9: Comparative organization of RGMc genes. The anatomy of human, mouse, and zebrafish RGMc genes is shown. Exons are indicated by *boxes*, with coding regions in *blue* and non-coding regions in *yellow*. The assignment of exon numbers is based on comparison with mouse RGMc, and is provisional for zebrafish (in parentheses). The polyadenylation site is represented by a *vertical arrow*.

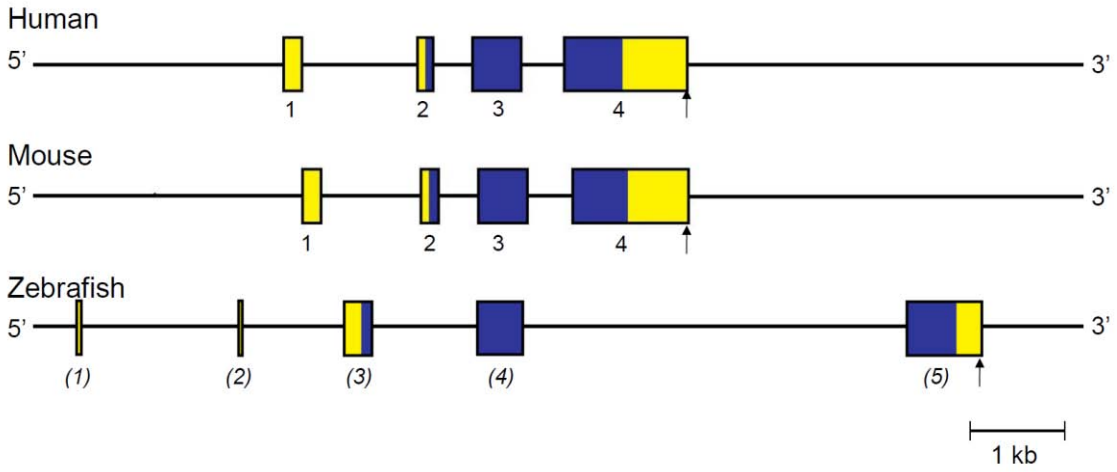

Human
5'
3'
1
2
3
4
Mouse
5'
3'
1
2
3
4
Zebrafish
5'
3'
(1)
(2)
(3)
(4)
(5)
1 kb

Figure 2.10: Map of the known mutations in Hemojuvelin (HJV; human RGMc) that cause the disease juvenile hemochromatosis (JH). The full length linear sequence of human Hemojuvelin (HJV; human RGMc), with the following annotations: SP (signal peptide) motif (light blue) and its approximate boundaries which have not been experimentally determined; RGD (arginine-glycine-aspartic acid) motif (green); vWD, partial vonWillebrand type-D domain (yellow); PPC, pro-protein convertase recognition and cleavage site (purple); squiggle at the COOH-terminus represents the GPI anchor signal sequence; locations of asparagine-linked glycosylation sites are indicated by asterisks, '*'; vertical solid arrowhead - site of intra-molecular proteolytic cleavage to generate two-chain HJV/RGMc. Above the linear sequence shown with a *bent arrow* are the known JH-causing frameshift (nonsense) mutations along with the amino acid residue at which the mutation takes place (e.g., L27fsX is a leucine at amino acid position 27). Below the sequence are the known JH-causing mis-sense mutation depicted as a *vertical line* along with the amino acid residue in the wild-type protein, the position in the linear sequence, and the residue to which the mutation creates (e.g., G320V (mouse G313V) is a glycine at amino acid position 320 mutated to a valine). Also see Table 2.8.

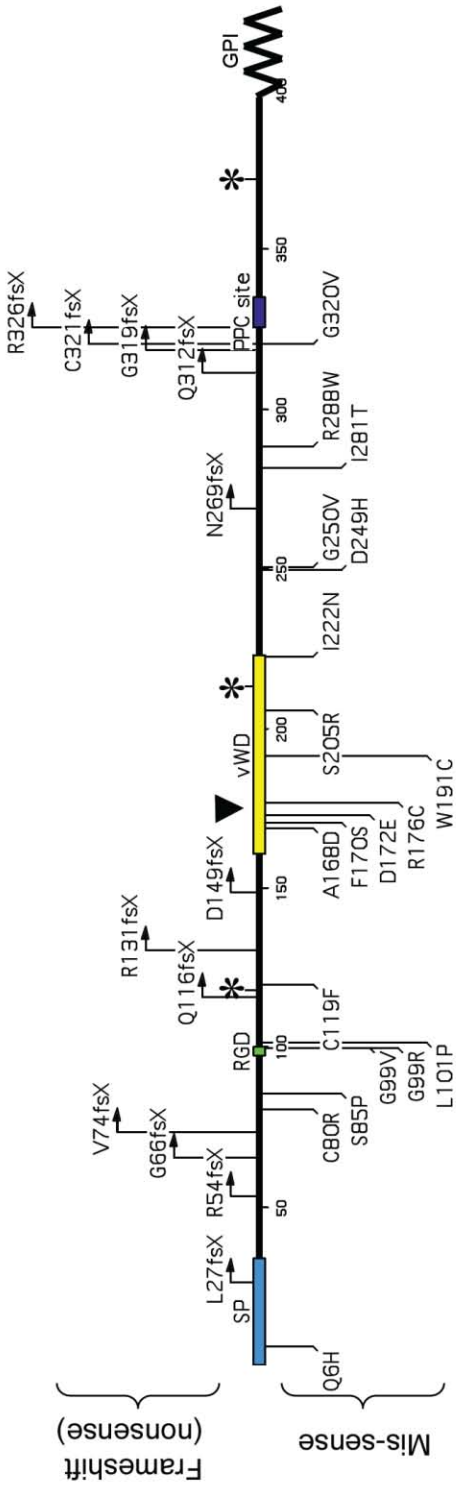

GPI
400
350
300
250
200
150
100
50
PPC site
vWD
SP
RGD
R326fsX
C321fsX
G319fsX
Q312fsX
G320V
R288W
I281T
N269fsX
G250V
D249H
I222N
S205R
W191C
A168D
F170S
D172E
R176C
D149fsX
R131fsX
Q116fsX
C119F
G99V
G99R
L101P
V74fsX
G66fsX
R54fsX
C80R
S85P
L27fsX
Q6H
Frameshift
(nonsense)
Mis-sense

Figure 2.11: Comparative organization of RGMc at exon boundaries within the coding sequence. The arrangement of the coding sequence (CDS) of RGMc appears to be well conserved over evolutionary distance. Human, mouse, and zebrafish RGMc all use a three-exon CDS in which the first and second coding exons share a codon triplet, whereas the second and third coding exons have separable codon triplets. RGMc genes from these three species are shown with exons indicated by *boxes*, with coding regions in *blue* and non-coding regions in *yellow*. Vertical lines with gaps indicate large distances in the zebrafish introns when compared to human and mouse; refer to Table 2.7 for intronic distances. On the bottom is the example of mouse (*Mmu = Mus musculus*) RGMc with the protein sequence enclosed in a box, and the corresponding nucleotide codon triplet indicated below in the same color (red or green). Exon boundaries are indicated by vertical lines.

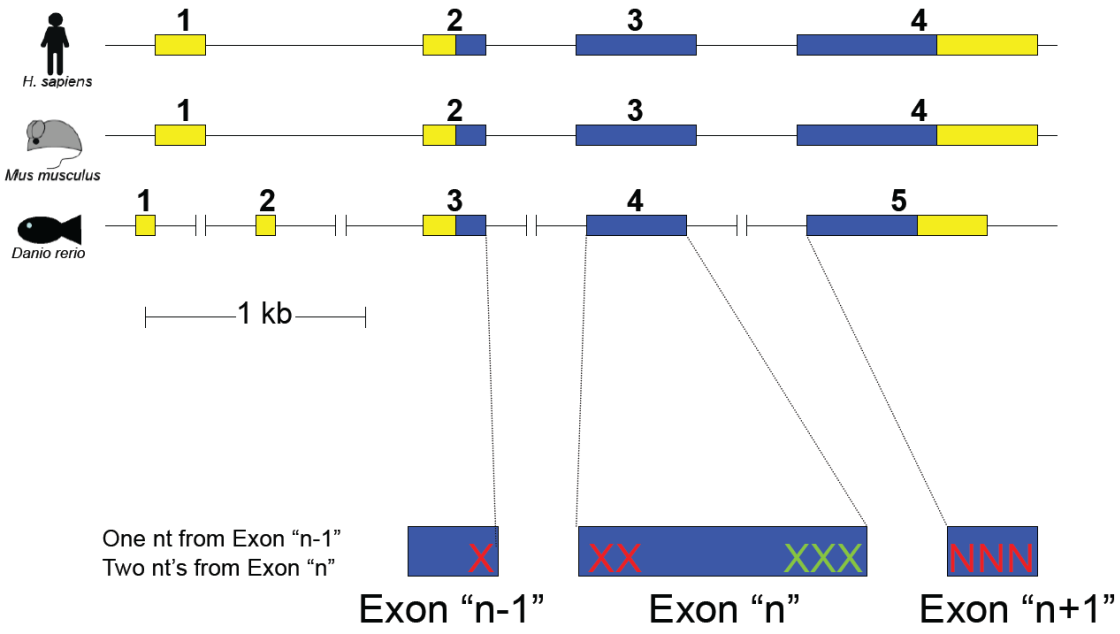

H. sapiens
1
2
3
4
Mus musculus
1
2
3
4
Danio rerio
1
2
3
4
5
1 kb
One nt from Exon “n-1”
Two nt’s from Exon “n”
X
XX
XXX
NNN
Exon “n-1”
Exon “n”
Exon “n+1”

Figure 2.12: Phylogeny of the RGM family. Evolutionary trees have been derived from the protein translation of well-annotated RGM DNA sequences in which the mRNA and gene agrees. Methods of analysis are as follows: 7 separate multiple sequence alignments (MSAs) of full-length RGM proteins were performed with MUSCLE [39], Clustal-W [114], or hand-alignment, followed by direct submission or a codon-optimized alignment through PAL2NAL [115]. Either protein-MSAs or codon-based alignments were submitted to several phylogenetic methods, including neighbor joining with unrooted and rooted trees (via MacVector), maximum likelihood [116, 117] (with and without Bootstrap methods on neighbor joining and maximum likelihood), and Bayesian [118, 119] analysis. **A**. RGM family phylogeny using an unrooted maximum likelihood method, displaying a distance of 0.1 amino acid substitutions per position (scale bar). **B**. RGM family cladogram derived from the neighbor joining method (Poisson-correction with gaps distributed proportionally) rooted with zebrafish (*Dre*) RGMc, displaying bootstrap values as percentage of 5000 replications supporting that branch on the cladogram. Species abbreviations for **A** and **B** may be found in Table 2.1. For both **A** and **B**, the putative ancestral RGM is highlighted in green and the ancestral gene to RGMa and RGMb is shown in blue. Phylogeny and cladogram created using Pal2NAL [115], Selection Server [120], Phylogeny.fr [117], PhyML 3.0 [116], TreeDyn [121], and MacVector v7.2.3.

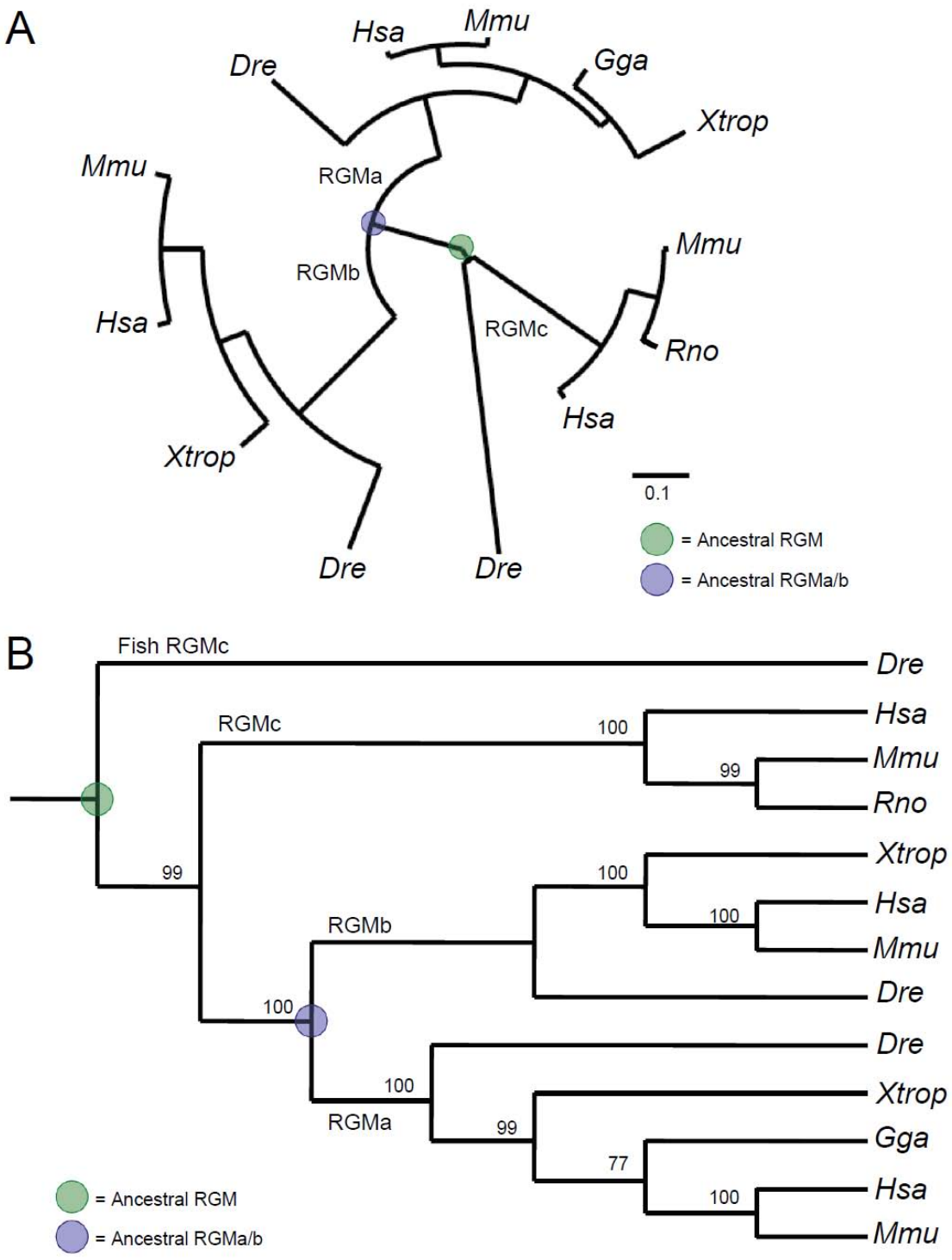
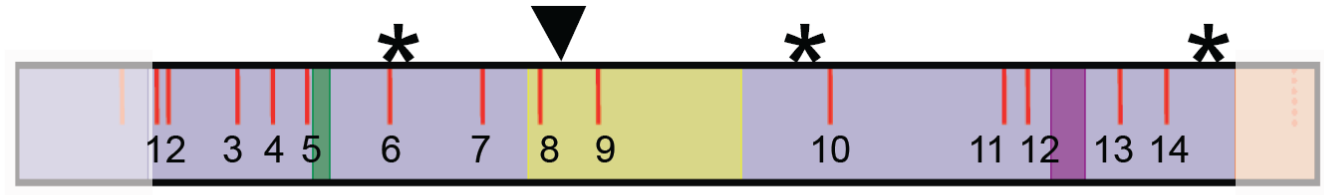

A
Hsa
Mmu
Dre
Gga
Xtrop
Mmu
RGMa
RGMb
Mmu
Hsa
RGMc
Rno
Hsa
Xtrop
0.1
Dre
Dre
= Ancestral RGM
= Ancestral RGMa/b
B
Fish RGMc
Dre
Hsa
RGMc
100
Mmu
99
Rno
Xtrop
99
100
Hsa
100
RGMb
Mmu
Dre
100
Dre
100
Xtrop
RGMa
99
Gga
77
Hsa
100
Mmu
= Ancestral RGM
= Ancestral RGMa/b

Figure 2.13: Mapping the putative disulfide bonds to the *Ab initio* model of the RGM proteins. The predicted disulfide bonding pattern of six independent *ab intio* models (see Fig. 2.14 for structure and details) are shown (e.g., Model 1: Cys^{1}-Cys^{2} are predicted to form a disulfide bond, as are Cys^{4}-Cys^{5}, Cys^{6}-Cys^{10}, etc.). Alternative bonding patterns within the same model are listed as 'alt.' if an additional cysteine was predicted to be in close proximity. Cysteines are shown as red lines, and numbered 1 to 14 according to the number found in the mature protein (all 14 cysteines in the mature protein are conserved in all species listed in Table 2.1). The signal peptide (SP) and GPI-anchor sequences are shaded, as they are not included in the mature protein, but present to demonstrate the cysteine found in the SP and numerous non-conserved cysteines that may be present in the GPI signal sequence of select RGM family members (depicted as a dotted red line). Additional regions of interest are labeled as follows: RGD motif (RGMa and RGMc only, *green*); vWD, partial vonWillebrand type-D domain (*yellow*); PPC, pro-protein convertase recognition and cleavage site (Mammalian RGMc only, *purple*); asparagine-linked glycosylation sites are indicated by asterisks, '*'; vertical solid arrowhead - site of intra-molecular proteolytic cleavage to generate the two-chain RGMa and RGMc.



Model:	1	2	3	4	5	6
	(1,2) (alt 1,4) (4,5) (6,10) (11,12) (13,14)	(3,5) (4,6) (7,9) (alt 1,9) (11,12) (13,14)	(1,2) (4,6) (alt 3,5) (7,9) (11,12) (13,14)	(1,2) (alt 1,7 1,3) (4,5) (alt 5,10) (8,9) (11,12) (13,14)	(1,5) (alt 1,4 4,5) (8,9) (11,12) (13,14)	(1,2) (alt 1,4 2,4) (4,5) (8,9) (alt 8,10) (11,12) (13,14)

Figure 2.14: A*b initio* model for RGM proteins. The model was generated using Rosetta [88-91, 94, 122], using the following steps: First, 1000 independent structures were predicted from a fragment library prepared with the Robetta Fragment server [87, 92, 93]. Structures were clustered for similarity based on their root mean square deviations. The centers of the three largest clusters were chosen as the best models, defined as having the lowest standard deviation of the mean among positions of carbon atoms of all residues to all other simulations in a cluster. Selected structures were minimized using CharmM [123, 124] and analyzed for consistency with known experimental data as described [125]. A single model is illustrated. **A**. Cartoon version of the model. Cylinders represent α-helical regions, thick lines with arrows, β-sheets, and thin lines, unstructured regions. The model suggests that members of the RGM-family adopt a two-lobe structure. The RGD domain is depicted in green, the partial vWD domain is in yellow, cysteines are in purple, asparagine-linked glycosylation sites conserved in all 3 mammalian RGMs are in cyan (and labeled –NCS– and –NFT–), and the GPI anchor attachment site at the COOH-terminus is noted. All of the above regions appear to be surface exposed. The pro-protein convertase site (found only in mammalian RGMc) is depicted by an arrow labeled PPC. The NH_2-terminus is not visible as it is located behind the partial vWD domain in the left lobe of the protein. An interactive three-dimensional version of **A** can be found at http://www.BiochemJ.org/422/0393/bj4220393add.him . **B**. Space-filling version of the model. The increasing thickness of the tubes represents greater divergence in primary amino acid sequences among RGM family members. The protein domains are color-coded as in **A**.

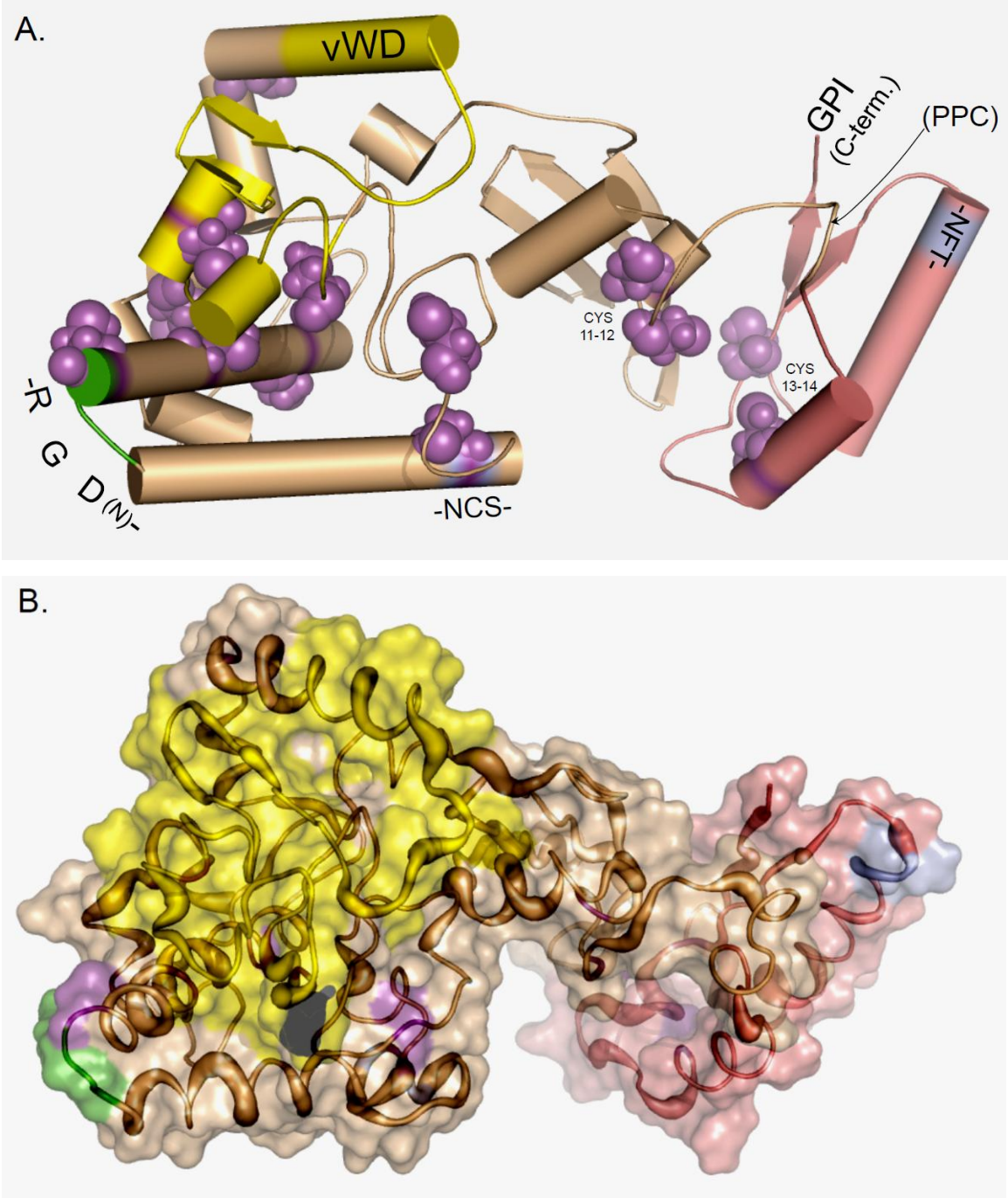
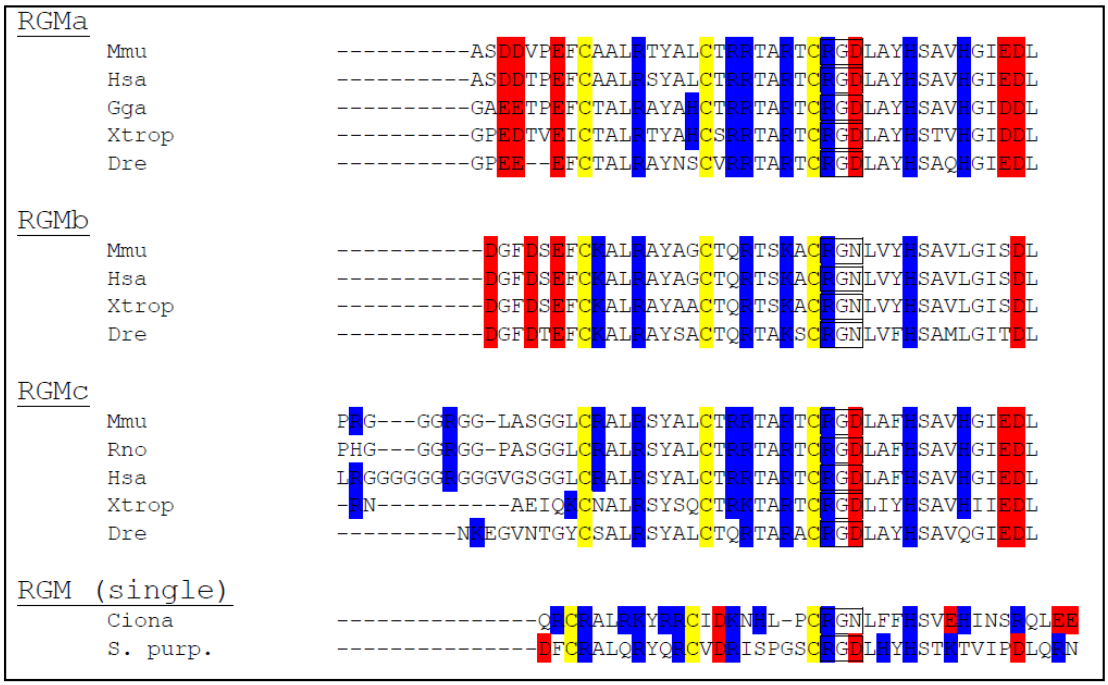
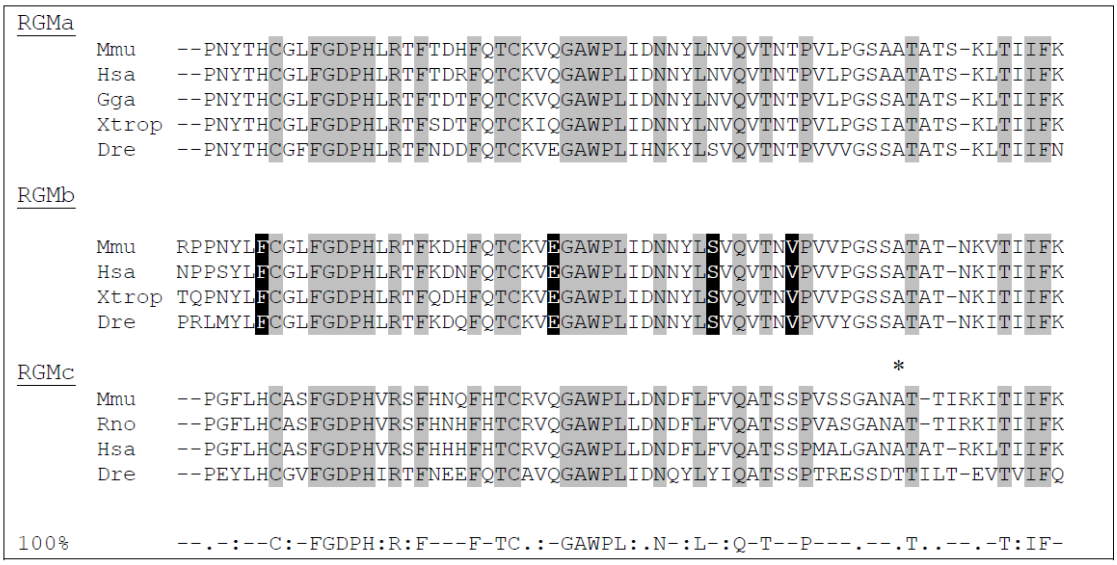

A.
vWD
GPI
(C-term.)
(PPC)
-NFT-
CYS
11-12
CYS
13-14
-R G D (N)-
-NCS-
B.

Figure 2.15: The RGD-motif in RGM proteins. Protein sequence alignment of the RGD (arginine-glycine-aspartic acid)-motif (found in RGMa, RGMc, and the putative RGM from purple sea urchin, *Strongylocentrotus purpuratus*) or –RGN– sequence (for RGMb and the putative RGM from *Ciona intestinalis*) highlighted with *boxes*. Cysteine residues highlighted in *yellow*, with acidic and basic residues highlighted in *red* and *blue*, respectively. Abbreviations for species listed may be found in Table 2.1.



Cys =
Acidic =
Basic =

Figure 2.16 The vWD-domain in RGM proteins. Protein sequence alignment of the partial vWD (vonWillebrand type-D)-domain. Regions conserved between all RGM family members are highlighted in *gray*, and areas where RGMb is highly conserved but differs from RGMa and RGMc are highlighted in *black.* The putative N-linked glycosylation site found only in mammalian RGMc is indicated by an asterisk, '*****'. Amino acid residues that are 100% conserved are shown at the bottom, along with the following notations: '**:**' conserved chemistry (e.g., hydrophobic), and '**.**' conserved in all but one sequence. Abbreviations for species listed may be found in Table 2.1.



: Conserved Chemistry (e.g., Hydrophobic)
. Conserved in all but one sequence

Areas conserved between all RGM family members =
Areas where RGMb differs from RGMa and RGMc =

(This page was intentionally left blank)

Chapter 3

Structure of the RGMc gene and characterization of the RGMc promoter

"What makes it difficult is that research is immersion in the unknown. We just don't know what we're doing. We can't be sure whether we're asking the right question or doing the right experiment until we get the answer or the result." –Martin A. Schwartz (J. Cell Science 2008)

"To arrive at knowledge slowly, by one's own experience, is better than to learn by rote in a hurry, the facts that other people know, and then be glutted with words, to lose one's own free, observant, and inquisitive ability to study." –Johann Heinrich Pestalozzi (1746-1827)

The majority of the research in this chapter has been submitted for publication as:

Conserved proximal promoter elements control repulsive guidance molecule c/hemojuvelin gene transcription

Christopher J. Severyn and Peter Rotwein

Department of Biochemistry and Molecular Biology, Oregon Health & Science University, 3181 SW Sam Jackson Park Road, Portland, OR 97239-3098, U.S.A.

Genomics (2010)

Received 12 June 2010

This work was supported by the National Institutes of Health, grant numbers R01 DK42748 (to P.R.), T32 HL007781 and F30 HL095327 (to C. J.S.).

3.1: Summary

Repulsive guidance molecule c (RGMc), or hemojuvelin (HJV), a protein first reported in 2004, is produced in striated muscle and the liver, has been shown to play a critical role in systemic iron metabolism in mammals. Inactivating mutations in RGMc/HJV cause juvenile hemochromatosis, a severe systemic iron overload disorder in humans. Understanding the molecular mechanisms responsible for control of RGMc biosynthesis under physiological or pathological conditions has been hampered by lack of fundamental information about the RGMc gene. In this study, we define the structure of the mouse RGMc gene and its mechanisms of regulation in skeletal muscle, the major tissue where it is expressed. RGMc is a 4-exon gene with a discrete transcription start site in exon 1 that undergoes alternative RNA splicing in exon 2 to yield 3 distinct mRNAs differing in length of the 5' untranslated region. RGMc gene transcription is induced early during myoblast differentiation in culture, leading to sustained production of RGMc mRNA. Using reporter gene experiments we identify 3 critical regions of the proximal RGMc gene promoter, comprising paired E-boxes, a putative Stat and/or Ets element, and a MEF2 site, that when mutated collectively abrogate RGMc transcriptional activity in muscle. Myogenin and MEF2C can activate the RGMc promoter in non-muscle cells, supporting the idea that these muscle-enriched transcription factors are key components of RGMc gene regulation. In contrast, we were not able to identify promoter elements responsible for RGMc gene expression in the liver. As the promoter elements are highly conserved in vertebrates, our results support the hypothesis that RGMc has been a muscle-enriched gene throughout its evolutionary history.

3.2: Introduction

Iron-related metabolic and hematologic disorders affect millions of individuals worldwide. Iron plays a critical role in numerous cellular processes ranging from oxygen exchange and energy metabolism [126] to nucleic acid synthesis and DNA repair [127], yet too much or too little iron can cause severe tissue and organ damage [24]. As a result, iron levels are tightly regulated in humans and other mammalian species [128], with primary control being exerted at the level of absorption from the small intestine [128, 129]. Hemojuvelin (HJV) or repulsive guidance molecule c (RGMc) is a recently identified gene that was initially linked to systemic iron metabolism by the discovery that mutations in humans caused the rapidly progressive and severe iron overload disorder, juvenile hemochromatosis (JH) [105, 130]. This relationship was strengthened when mice engineered to lack RGMc also developed iron overload [1, 2]. As RGMc/HJV has been found more recently to indirectly regulate the expression of hepcidin [3, 69, 72], a peptide hormone made in the liver that negatively controls intestinal iron absorption in the duodenum [74, 129], it is thus a component of a homeostatic pathway that regulates iron uptake [1-3, 24, 72].

RGMc was discovered not only as a gene mutated in JH [3], but also was identified as a novel transcript expressed during skeletal muscle differentiation [31], and as a member of a conserved three-gene family that receives its name from the axonal guidance molecule RGMa [27, 29, 30, 33]. Unlike RGMa or RGMb, RGMc is not expressed in the central nervous system, but rather is produced by striated muscle and by hepatocytes in the liver [27, 31, 33]. During development RGMc transcripts are expressed first in the somites in both mice and zebrafish [30, 31, 41] (Fig. 3.1A and B), and then later in skeletal muscle, as well as in the embryonic heart and liver [1, 31] (Fig. 3.1A). This unique pattern of

RGMc expression in striated myocytes and hepatocytes is maintained in the adult (Fig. 3.1C). To date, the molecular mechanisms responsible for tissue-specific gene expression have not been elucidated, and very little is known about RGMc gene regulation in any species, as no promoter has been characterized. Here we define the structure of the mouse RGMc gene and identify the DNA elements responsible for gene transcription in skeletal muscle. Further analysis reveals that these *cis*-acting muscle-specifying DNA elements are highly conserved in RGMc genes from multiple mammalian species, supporting the hypothesis that RGMc has been a muscle-enriched gene throughout its evolutionary history. Furthermore, many of these elements are conserved in non-mammalian vertebrates and can be found in other RGM gene family members. I will explore the implications of these conserved elements to gain future insight into the evolution of the RGM gene familiy, while focusing experimentally on the regulatory mechanisms of the RGMc gene.

3.3: Experimental Procedures

3.3.1: Materials – Restriction enzymes, buffers, ligases, and polymerases were purchased from New England Biolabs (Beverly, MA), BD Biosciences (Clontech, Mountain View, CA), and Fermentas (Hanover, MD). The BCA protein assay kit was from Pierce (Thermo Scientific Life Sciences, Rockford, IL), and the QuikChange XL site-directed mutagenesis kit from Stratagene (La Jolla, CA). TransIT LT-1 was from Mirus Corp (Madison, WI). Dulbecco's modified Eagle's medium (DMEM), Superscript III first-strand synthesis kit, Trizol, and horse serum were from Life Technologies (Carlsbad, CA). Fetal bovine serum (FBS) and newborn calf serum (NCS) were from Hyclone (Logan, UT). Luciferase assay reagent was purchased from Promega (Madison, WI). AquaBlock EIA/WIB solution was from East Coast Biologicals (North Berwick, ME), and Immobilon-FL was from Millipore (Billerico, MA). DRB (5,6-dichloro-1-β-D-ribofuranosylbenzimidazole) was from Sigma (St. Louis, MO), and DNA purification reagents from Qiagen (Valencia, CA). Antibodies were purchased from the following suppliers: myogenin (F5D hybridoma), Developmental Studies Hybridoma Bank (Iowa City, IA; W. E. Wright); MEF2 (C-21, sc-313), and MyoD (M-318, sc-760), Santa Cruz Biotechnologies (Santa Cruz, CA); α-tubulin, Sigma; Pan Akt, (#9272), Cell Signaling Technology (Beverly, MA); CREB (UBI 06-863), Upstate Biochemicals, a division of Millipore; AlexaFluor 680- and 800-conjugated goat anti-mouse IgG, Life Technologies; IR800-conjugated goat anti-rabbit IgG, Rockland (Gilbertsville, PA). Oligonucleotides were synthesized at the OHSU DNA Services Core. All other chemicals were reagent grade and were purchased from commercial suppliers.

3.3.2: Construction of RGMc promoter-reporter plasmids – Mouse RGMc genomic DNA was isolated from BAC clone RP24-136I19 (Children's Hospital Oakland Research

Institute BACPAC resource center (http://bacpac.chori.org/), Oakland, CA). All RGMc DNA sequences characterized in this dissertation match what is present in mouse genome databases [45] except for a 33-bp region of intron 2, which is not present in the BAC DNA. DNA fragments generated by restriction enzyme digestion or PCR were purified after preparative agarose gel electrophoresis by ion-exchange chromatography (Qiaexx II gel extraction kit, Qiagen), and sub-cloned into the pGL3-basic firefly luciferase vector (Promega) by standard methods. Mutations in the RGMc promoter were introduced by site-directed mutagenesis with the following oligonucleotide primers (top strand is shown, mutations are in *lower case*):

E-box at -588 to -583:

5'-GGTGGAGAGAGAGTAGAgctagcCAGAGATCTGATCTGGGC-3';

E-box at -514 to -509:

5'-GCTCTCGGATTTCTCGGGAgggcccGACCTTTCAGCTTCTG-3';

β-site at -119 to -111:

5'-GCTCCCACACCCCACTGCCACCAACGcgtCTGcccTTTTGGACCTAG-3';

MEF2 site at -98 to -84:

5'-CCTGGAATTTTGGACCTAGtggccaTTAgAAtTcTCAACTCAGTAGGC ACCTCCCTCCTCC-3'.

All DNA modifications were confirmed by sequencing.

3.3.3: Cell culture - Cells were incubated at 37°C in humidified air with 5% CO_2. C2 myoblasts (passages 4 to 10) were grown on gelatin-coated tissue culture dishes in DMEM with 10% heat-inactivated FBS and 10% NCS. At confluent density cells were washed and low-serum differentiation medium was added (DM = DMEM with 2% horse serum). C3H-10T½ mouse embryonic fibroblasts (10T½ cells, #CCL-226, ATCC, Manassas, VA) were maintained between passages 12 and 18, and incubated on gelatin-

coated dishes in DMEM plus 10% FBS. They were converted to myoblasts by infection with a recombinant adenovirus for mouse MyoD [131], followed by incubation in DM as above. Liver cell cultures were as follows: Normal alpha mouse liver 12 (AML-12; CRL-2254) cells, cultured in a 1:1 mixture of Dulbecco's modified Eagle's medium (DMEM) and Ham's F12 medium supplemented with ITS (Final concentration:10 μg/ml insulin, 5.5 μg/ml transferrin, 6.7 ng/ml selenium), 0.1 μM dexamethasone, and 10% FBS. Two different human hepatoma cell lines were used (*HepG2*, HB-8065; and *Hep3B*, HB-8064), and were cultured in Eagle's Minimum Essential Medium (EMEM) plus 10% FBS. Primary mouse hepatocytes were harvested by A. Duncan of the Grompe Laboratory at OHSU via methods described in Refs. [132-134], and subsequently cultured by the author as per above references, outlined in Fig. 3.19A. Rat hepatoma FTO-2B cells were cultured in DMEM plus 10% FBS, and kindly provided by M. Thayer of OHSU. Cell images were captured by phase contrast microscopy using a Nikon Eclipse T300 microscope with an attached Roper Scientific Cool Snap FX CCD camera.

3.3.4: Recombinant adenoviruses - Recombinant adenoviruses for MyoD (Ad-MyoD) and β-galactosidase (Ad-β-gal) were prepared as described [131].

3.3.5: Analysis of gene transcription by promoter-reporter gene assays - C2 and 10T½ on gelatin-coated 12-well plates were transfected with individual RGMc promoter-reporter plasmids or with controls [(thymidine-kinase (TK) - Luc [135], mouse myogenin - Luc [136], mouse IGF-II promoter 3 - Luc [137], 4xE-box TK - Luc [136]] at 50% or 25% of confluent density, respectively (0.7 μg of plasmid DNA per well for C2 cells, 0.4 μg for 10T½ cells). C2 cell extracts were harvested either one-day later (undifferentiated), or after DM was added for an additional 48 hr (differentiated). For

10T½ cells, one day after transfection cells were infected with Ad-MyoD or Ad-β-gal, and extracts were collected following another day in growth medium (undifferentiated), or after DM was added for a further 24 hr (differentiated). Cell extracts from individual experiments were stored at -80^{o}C, and were assayed together for luciferase activity, as described [136], and results were normalized to cellular protein concentrations. At least 3 experiments were performed for each promoter - reporter plasmid using duplicate transfections per experiment. To assess effects of myogenin or MEF2 on RGMc promoter activity, co-transfection experiments were performed in 10T½ cells with selected RGMc promoter - reporter genes, and expression plasmids for mouse myogenin (myogenin-IRES-EGFP or EGFP [136]), or constitutively active MEF2C (MEF2C-VP16 - from J. Molkentin [138, 139]) in pcDNA3 or empty vector.

3.3.6: Animal studies - Male C57Bl6 mice were housed at the OHSU Animal Care Facility on a 12 hr light/dark schedule with free access to food and water, and received care according to National Institutes of Health guidelines. At 3 months of age, mice were euthanized by cervical dislocation and tissues were harvested, flash-frozen in liquid nitrogen, and pulverized prior to RNA isolation. Animal studies were approved by the OHSU Animal Care and Use Committee.

3.3.7: RNA isolation and analysis - Total cellular RNA was isolated from cells and tissues using Trizol, followed by sodium acetate-ethanol precipitation and suspension in RNAse-free de-ionized water. Nuclear RNA was isolated from cells as described [137]. RNA concentrations were determined spectrophotometrically at 260 nm, and quality assessed by agarose gel electrophoresis. RNA (5 µg) was reverse transcribed in a final volume of 20 µl, with either oligo-dT primers (for total RNA) or random hexamers (for

nuclear RNA), and PCR was performed with 0.1 µl of cDNA and the primers listed in Table 3.1. The linear range of product amplification was established in pilot studies for each primer pair, and the cycle number that reflected the approximate midpoint was used in final experiments. This varied from 18 - 30 cycles for total RNA and from 25 - 30 cycles for nuclear RNA. Results were visualized after electrophoresis through 1.0 - 1.8% agarose or 10% PAGE gels.

3.3.8: RNA half-life – Confluent C2 cells were incubated in DM for 48 hr, washed, and DRB [75 µM] was added in DM for 0.5 to 12 hr. Total cellular RNA was isolated, and used in RT-PCR experiments, as above. Half-life was determined by averaging results of two independent experiments using non-linear regression fit to a *one-phase decay* equation, $Y = Y_0 \bullet e^{-kX}$.

3.3.9: Mapping the 5' end of the mouse RGMc gene— The 5' RACE method was employed with mouse skeletal muscle RNA. First strand cDNA was prepared using specific primers complementary to portions of mouse RGMc exons 2 and 4 (Table 3.1). Subsequent steps were as described [140, 141], with primers for second strand cDNA synthesis from RGMc exon 1 (Table 3.1), and for PCR from RGMc exon 1 and a poly T adaptor (Table 3.1). Gel-purified PCR products were cloned into the pGEM-T Easy vector (Promega), and the DNA was sequenced. A total of 15 independent clones were analyzed. A nested RT-PCR-based method also was used to map the 5' extent of RGMc exon 1 from mouse muscle, liver, and heart RNA. Primers are listed in Table 3.1. A total of 83 independent clones were characterized by restriction enzyme mapping and/or DNA sequencing.

3.3.10: DNA-Protein binding studies— Electrophoretic mobility shift assays were performed as described in [135, 142, 143], with C2 or 10T½ nuclear proteins and 5'-phoshphoramidite-infrared-dye700-labeled double-stranded oligonucleotides as follows (only the top strand is listed):

Gene from:	Oligo name	Length	DNA Seq. (Top Strand)
mMCK	MEF2 (E. Olsen) Ref.[144]	25 nt	5' GATCGCTCTAAAAATAACCCTGTGC
	MEF2 (mMCK)	36 nt	5' GGAGGAGAAGCTCGCTCTAAAAATAACCCTGTCCCT
RGMc	RGMc (-106:-80)	27 nt	5' GGACCTAGCTATTTTTAAAACTGTCAA
	RGMc (-127:-98)	30 nt	5' CACCAACGTTCCTGGAATTTTGGACCTAGC
	RGMc (-143:-110)	34 nt	5' TCCCACACCCCACTGCCACCAACGTTCCTGGAAT
	RGMc (+116:146)	30 nt	5' GGCTCGAGAACCCAGTATCAGAGTAATGCT
	RGMc (+136:165)	30 nt	5' GAGTAATGCTTGACCTCGGGAAACAGTAAG
RGMc MEF2mut	RGMc-MEF2mut (-106:-80)	27 nt	5' GGACCTAGC***GGGG***TTTAAAACTGTCAA

After incubation of proteins and DNA for 30 min at 4°C, products were separated by electrophoresis through non-denaturing 5% polyacrylamide gels in TBE (90mM Tris, 90mM boric acid, and 2 mM EDTA, pH 8.3) at 200 V for 25–35 min at 4 °C. Results were analyzed using the LiCor-Odyssey infrared imaging system and v1.2 analysis software (LiCoR Biosciences, Lincoln, NB). The top strand of the nonspecific competitor oligonucleotide is as follows: Oct-1: 5'TTTTAGAGGATCCATGCAAATGGACGTACG.

3.3.11: Protein isolation and immunoblotting - Whole cell protein lysates were prepared as described [131] and aliquots stored at -80°C until use. Protein samples (25 μg/lane) were separated by SDS-PAGE, transferred to Immobilon-FL, blocked in AquaBlock, and incubated with primary and secondary antibodies at the following dilutions: anti-myogenin (1:200), anti-MEF2 (1:3000), anti-MyoD (1:1000), anti-Akt (1:1000), anti-CREB (1:1000), anti-α-tubulin (1:30,000), and AlexaFluor 680-conjugated

goat anti-mouse IgG or IR800-conjugated goat anti-rabbit IgG (1:5000). Immunoblot images were acquired using a LiCoR Odyssey Infrared Imaging System, and analyzed with v2.0 analysis software (LiCoR, Lincoln, NE).

3.3.12: Data Analysis - Results were graphed and analyzed using Prism (GraphPad Software, San Diego, CA) or Excel (Microsoft, Redmond, WA). All differences were assessed using Student's *t-test* with $p < 0.05$ as a cut-off for significance.

3.3.13: Computational (in silico) Resources— UCSC Genome Browser [45], LLNL-ECR Browser [145], Dcode [146], rVista 2.0 [147], TRANSFAC database [148], GenBank [149], JASPAR [150], CpG Plot [151]

3.4: Results

3.4.1: Defining RGMc gene structure.

RGMc appears to be a 4-exon gene in mice and humans [152], but the 5' end of exon 1 has not been established in either species, and the promoter has not been characterized. We thus mapped the start site for mouse RGMc gene transcription as a means to first identify and then analyze the promoter. RGMc mRNA is expressed in adult mouse skeletal and cardiac muscle and in the liver (Fig. 3.1C), as during development [1, 31] (Fig. 3.1). We used 5' RACE to determine the RGMc transcription start site (TSS) in skeletal muscle, and found that the 5' end of 14/15 independent cDNA clones clustered within a 5-nucleotide region of genomic DNA that mapped ~25 nucleotides downstream from a putative TATA box [153] (Fig. 3.2A). We also used RT-PCR with overlapping exon 1-specific primers to map the 5' extent of RGMc exon 1 from mouse muscle RNA, and obtained results in agreement with the 5' RACE data (Fig. 3.2B). Similar observations were seen with RNA from mouse heart and liver (Fig. 3.3A).

The RT-PCR experiments designed to map the 5' end of RGMc exon 1 used a common primer located in exon 2 (Fig. 3.2B), and results consistently yielded 3 distinct cDNAs that differed in length by 18 to 77 nucleotides (Fig. 3.2B and Fig. 3.3). By DNA sequencing, all three classes of cDNAs contained RGMc exons 1 and 2, and matched mouse RGMc genomic DNA (Fig. 3.2C and Fig. 3.3). These results support the idea that the RGMc gene undergoes alternative RNA splicing to generate transcripts with varying lengths of exon 2, a hypothesis confirmed by evidence for splice acceptor sites at each of the three putative junctions between intron 1 and exon 2 (AG nucleotides underlined in Fig. 3.2C). Additional support comes from an expressed sequence tag in GenBank (AI196626), which matches the intermediate sized version of mouse exon 2, and from

comparative genomic analysis of 10 mammalian species, which reveals sufficiently similar DNA sequences to indicate that alternative RNA splicing may be a common feature of RGMc genes (Fig. 3.3B). Furthermore, this variable-length 5'untranslated region (UTR) places RGMc transcripts above the average length for UTRs [154-157] (Fig. 3.4), which are often associated with changes in translatability of the mRNA [22, 158]. Additional analysis and information about the possible function of this alternative RNA splicing in RGMc may be found in the 'discussion section' of chapter 4. Taken together, our results show that mouse RGMc is a 4-exon gene with a discrete transcription start site in exon 1 and alternative RNA splicing involving exon 2 that leads to three distinct transcripts that vary in the length of the 5' UTR (Fig. 3.2D).

3.4.2: RGMc gene transcription is induced during skeletal muscle differentiation.

We next examined the kinetics of RGMc gene expression in differentiating skeletal muscle cells, using the well-characterized C2 myoblast line as a model [159-161] (Fig. 3.5A). RGMc mRNA was detected as early as 12 hr after addition of DM to confluent C2 cells, and its abundance increased progressively during the subsequent 60 hr in a pattern similar to myogenin, a critical transcription factor that is expressed early in muscle differentiation [162, 163] (Fig. 3.5B). Accumulation of RGMc mRNA during muscle differentiation appeared to be a consequence of induction of RGMc gene transcription, as measured by stimulation of nascent nuclear RGMc RNA beginning at ~8 hr after addition of DM, a pattern also temporally similar to myogenin (Fig. 3.5D). We also examined the RGMc mRNA stability beginning after 48 hr of differentiation by using the transcriptional elongation inhibitor, DRB [164], and found that RGMc appears to be a moderately long-lived mRNA, with a half-life of ~5.2 hrs, more than twice that of myogenin, and nearly four times as long as MyoD (Fig. 3.6). Taken together, results in

figures 3.5 and 3.6 demonstrate that induction of RGMc gene transcription is the major regulatory step responsible for accumulation of RGMc mRNA in differentiating myoblasts.

3.4.3: Analysis of RGMc promoter function in muscle differentiation.

To investigate RGMc promoter function, we first used a 4.2 kb DNA fragment located 5' to exon 1 to drive luciferase reporter activity in 10T½ cells infected with Ad-MyoD [136] (characterized in Fig. 3.7) and in C2 myoblasts. This genomic DNA fragment was isolated as it contained, and extended beyond, a well-conserved region between mouse and human (Fig. 3.8). In each cell line, RGMc promoter activity was enhanced by ~35-fold after induction of differentiation to levels ~10% of a myogenin promoter - reporter plasmid, whose activity also was stimulated during differentiation (Fig. 3.9). In contrast, neither promoter was active in 10T½ cells infected with Ad-β-gal and incubated in DM (data not shown), providing further support for the idea that RGMc transcriptional activity is up-regulated during muscle differentiation.

Mapping experiments were next performed with a series of 5' promoter deletions to define the DNA segments responsible for the transcriptional activity of RGMc during muscle differentiation. Three major regions were identified in Ad-MyoD-infected 10T½ myoblasts, based on a decline in luciferase activity when each segment was eliminated: nucleotides -620 to -506, -136 to -110, and -110 to -88 (Fig. 3.10). Similar results were observed in C2 cells, although the overall signal was lower (Fig. 3.10). In addition, more detailed mapping of the region revealed similar trends (Fig. 3.11).

As transcriptional regulatory elements may be found anywhere surrounding a gene, not just 5' to the TSS, we tested other parts of the RGMc gene for the presence of additional transcriptional control regions by cloning three genomic segments comprising ~ 7.5 kb of RGMc chromosomal DNA into the RGMc promoter (-620 to +118) - luciferase reporter plasmid (Fig. 3.12). In particular, a region of ~150 bps was well-conserved between mouse and human (Fig. 3.12A), suggesting a possible novel regulatory element downstream of the RGMc transcribed region. However, none of these DNA fragments altered RGMc promoter activity prior to or after onset of muscle differentiation in Ad-MyoD-infected 10T½ cells (Fig. 3.12C), indicating that no transcriptional enhancers or repressors were located within or immediately 3' to the RGMc structural gene. Collectively, this suggests that within an ~12 kb locus surrounding the ~4kb mouse RGMc gene, the major transcriptional regulatory elements are found within a region smaller than a kilobase (Figs. 3.10-3.12).

3.4.4: Identification of proximal promoter elements responsible for RGMc transcriptional activity during muscle differentiation.

We introduced nucleotide substitutions into each of the DNA elements in the RGMc promoter found to be functionally important for transcriptional activity during muscle differentiation (Fig. 3.10). Additionally these mutations are based on conservation among mammalian RGMc (Fig. 3.14) and sequences that represent putative transcription factor binding sites from several computational algorithms (See section 3.3.13, Refs. [145-148, 150]). Mutation of two putative E-boxes in the segment from -620 to -506 (α-element, Figs. 3.13A and 3.14) resulted in a 50% decrease in luciferase activity in differentiating Ad-MyoD 10T½ cells, and a 25% decline in C2 myoblasts (Fig. 3.13B), while disruption of a putative MEF2 site from -110 to -88, (γ-element, Fig. 3.13) caused a

50% reduction in Ad-MyoD 10T½ cells, and a ~75% decrease in C2 myoblasts (Fig. 3.13B). By contrast, elimination of a potential Stat binding site [TTCN$_3$GAA [165-167]] and/or Ets element [GGA(A/T) [168-170]] in the segment from -136 to -110 (β-element, Fig. 3.13), was less effective, and led to only a ~25% decline in reporter gene expression in Ad-MyoD 10T½ cells, and had no effect in C2 myoblasts, although when combined with mutation of the γ region, RGMc promoter activity was decreased by ~90% in C2 cells (Fig. 3.13). Mutation of all three elements reduced reporter gene expression to basal levels in both muscle cell lines, confirming that together these three sites are responsible for the entire induction of RGMc promoter activity seen during skeletal muscle differentiation. Comparison of sequences from several mammalian genomes reveals that the E-boxes and MEF2 sites are well-conserved among mammalian vertebrates, and that the β–site, a putative Stat and/or Ets site, is less well-conserved but still present in most mammals (Fig. 3.14).

As the γ–element/MEF2-site appears to be a major regulatory element, we sought to determine if it was physically bound by a MEF2 protein, as well as establish if the neighboring β-site was bound by a transcription factor. The electrophoretic mobility shift assay (EMSA, or 'gel-shift') was employed in order to determine the protein-DNA interactions at these regions. A series of three overlapping, double-stranded oligonucleotides ('oligos') with a 5'-infrared (IR) phosphoramidite-label (Fig. 3.15A) were generated, along with high quality nuclear extracts (Fig. 3.15B, *right*). As a control we turned to the well-characterized MEF2 site in the promoter of the mouse muscle creatine kinase (mMCK) gene [171-174]. MEF2-IR-oligos incubated with nuclear protein extracts from C2 cells (at t0 and t48 hrs following addition of DM; Fig. 3.15C *left panel, lanes3 and 6*, respectively) and 10T½ fibroblasts (*lane 8, left panel*) revealed a gel-shift pattern similar to seen previously with the mMCK MEF2 site [171, 174]. This

distinct doublet became more pronounced in C2 cells that have undergone differentiation (*compare lanes 3 and 6, left panel* Fig. 3.15C*)*, and may be competed off using an unlabeled ('cold') MEF2-oligo using 1x (*lane 4*) or 10x (*lanes 5, 7, and 9*) molar ratio of unlabeled to labeled oligo-probe. Furthermore, the gel-shift pattern with C2 nuclear extract alone (Fig. 3.15C, *right panel, lane 3*) is abolished with 10x unlabeled MEF2 (*lane 4, right panel*), but not with 10x molar ratio of a non-specific probe to Oct-1 (Fig. 3.15C, *right panel, lane 5*). When compared to a probe for the γ–element of RGMc (Fig. 3.15A, double-stranded oligo with '*'), a similar banding patter was seen with nuclear extracts from differentiated C2 cells (Fig. 3.15D, *lane 2*) as was found in the MEF2 site from the mMCK promoter (Fig. 3.15D, compare the RGMc probe in *lane 2* to *lanes 3-7*, mMCK). This suggests that the γ–element of RGMc promoter is physically bound by a member of the MEF2-family of transcription factors under conditions of muscle differentiation, the time when RGMc is expressed (Fig. 3.5). In addition, conservation of this element in other mammals (see Fig. 3.14) is highly suggestive that this element has been important for the regulation of RGMc throughout the evolution of the gene, at least in mammalian vertebrates (more about possible implications to RGMc in non-mammalian vertebrates and other RGM family members discussed below). As binding with other probes from the RGMc promoter was extremely weak and barely detectable, we decided to utilize an overexpression system coupled to a promoter-reporter assay to determine the effects of different transcription factor proteins at all three elements in the RGMc promoter.

3.4.5: Over-expressed MEF2 and myogenin stimulate RGMc promoter activity in mesenchymal stem cells.

Genes expressed in muscle are often regulated by transcription factors of the myogenic basic helix-loop-helix (bHLH), as well as members of the MEF2-family of proteins. The

bHLH transcription factors of the MyoD family (MyoD, myogenin, Myf5, and Mrf4) act as central regulators in genes expressed during the formation of muscle in cell culture and *in vivo* [175-177]. MyoD induces expression of the MEF2-family of transcription factors, which synergize with MyoD to activate many muscle-specific genes. While MEF2 has been shown to be important for the regulation of many muscle genes, it has also been shown to control gene expression in non-muscle cells, where it collaborates with other transcription factors [16]. Thus, the genes that are activated by MEF2 in different cell types depend largely on the extracellular signaling and interactions with co-factors [16, 138, 178, 179]. To directly test the role of MEF2 proteins and myogenic bHLH transcription factors in regulating RGMc promoter function, we performed co-transfection experiments in the 10T½ mesenchymal stem cell line (Fig. 3.16A). Constitutively active MEF2C (MEF2C-VP16) was able to boost the activity of the wild-type RGMc promoter by ~10-fold compared with the empty vector control, and this stimulation was lost after mutation of the MEF2 site (Fig. 3.16D). Myogenin was able to increase wild-type RGMc promoter activity by > 20-fold compared to a vector control, but this was reduced by only ~50% for the α-element mutant (Fig. 3.16B), possibly because another conserved E-box is found at -53 to -48 (see Fig. 3.14). In contrast to these observations, activity of the wild-type RGMc promoter was not induced by an expression plasmid for constitutively active Stat5b (3.16C). Collectively, these results suggest that the RGMc promoter in skeletal muscle is regulated by three elements, two of which are regulated by a member of the bHLH family (e.g., myogenin) and MEF2 (e.g., MEF2C) at the α- and γ-sites, respectively (summarized in Fig. 3.17, discussed in detail below).

3.4.6: Liver-specific control elements in the RGMc promoter.

Regulation of genes expressed in the liver follow similar fundamental principles seen in

muscle, but the transcriptional regulatory factors that bind are often unique to, or highly-enriched, in the liver. While RGMc is most highly expressed in striated muscle (skeletal and cardiac) and to a lesser degree in the liver ([1, 27, 29, 31, 33, 34] and Fig. 3.1), the liver is currently recognized as an important mediator of iron homeostasis. Therefore, we sought to determine the regulatory mechanisms that control RGMc expression in the liver in addition to muscle. Mouse studies using an RGMc-targeted interruption knockout (containing an expression marker) demonstrated that reporter expression appears to be selectively localized to the periportal hepatocytes [1]. Developmentally, RGMc expression occurs in the embryonic liver of mice at around 12.5dpc (days post conception) [180]. Interestingly, the important γ-element for RGMc expression in muscle (a highly conserved MEF2 site) overlaps with a consensus site for the liver-enriched transcription factor, hepatocyte nuclear factor (HNF)-4α, an orphan member of the nuclear receptor family [181] with a paired zinc-finger domain [182]. More importantly, HNF4α has been shown to be localized at two regions in the human HJV/RGMc locus via chromatin immunoprecipitation followed by DNA sequencing (ChIP-Seq) data (personal communications with, and data acquired by the author from M.D. Wilson of the Odom laboratory, published in Ref. [183]). Furthermore, detailed analysis by the author of microarray studies of a mouse knockout (KO) for HNF4α suggests that RGMc fails to be expressed in the knockout at any time during development [184]. Given this preliminary data, I would postulate that this HNF4α consensus site, just like the MEF2/ γ-element in muscle, is a critical location for transcription factor binding, co-factor recruitment, and subsequent expression of RGMc in the liver. Unfortunately, to date there have been no examples of endogenous RGMc expression in any cell line. Furthermore, my preliminary studies demonstrate that there is no RGMc reporter gene activity in three unique cell lines tested (Fig. 3.18, detailed below).

The same promoter-reporter constructs used in the experiments to define regions that

control RGMc expression in muscle were transfected into three different liver cell lines, normal mouse liver cells (AML-12), and two human hepatoma cell lines (HepG2 and Hep3B) (Fig. 3.18). As shown in figure 3.18, none of the RGMc promoter-reporter constructs showed activity in AML-12, HepG2, and Hep3B cells, although mouse insulin-like growth factor-II (IGF-II) promoter 3 had robust expression, and a neutral thymidine kinase (TK)-luciferase promoter had a basal level of activity (Fig. 3.18B). Since RGMc mRNA could not be detected in these liver cell lines (data not shown), but is present in adult liver (Fig.3.1C and Ref.[1]), I hypothesize that transcription factors critical for RGMc gene expression are not provided by these cell lines. The leading candidate, based on ChIP-Seq and microarray expression data from knockout mice (data analysis of Refs. [183, 184]), is the liver-enriched transcription factor HNF4α.

It has been well-documented that liver cells in primary culture undergo an adaptation in which the cells appear to "de-differentiate" and lose expression of many genes normally expressed *in vivo*. Thus it is not entirely surprising that none of the liver cell lines used in figure 3.18 show activity with the RGMc promoter constructs. For example, Clayton and Darnell demonstrated that the ^{32}P-labeled mRNAs dramatically decrease over the first 24 hours (e.g., albumin: 3-fold within the first 7 hours and >50-fold within 24 hours), but that this time varied depending on the gene; however, most genes decreased expression gradually over 140 hours [185]. To overcome this challenge I turned to primary hepatocytes that were isolated and cultured as previously described [132-134] (overview in Fig. 3.19). By isolating mRNA from primary hepatocytes in culture at intervals over several days, followed by RT-PCR, I was able to determine when the steady-state levels of RGMc began to decrease, and therefore empirically determine the "window of time" for transfection of RGMc reporter constructs. As shown in figure 3.19C, the decrease in RGMc expression in hepatocyte primary cell cultures is quite rapid with less than half of the starting levels of mRNA by 28 hours in culture, and virtually none by 148 hours. The

kinetics of RGMc mRNA expression suggest that future experiments of transfecting primary hepatocytes with RGMc promoter-reporter plasmids to map the functional elements responsible for regulating RGMc gene transcription in the liver may be quite challenging. Alternatively and independent of expression studies, DNA-protein experiments using hepatocyte nuclear protein extracts for footprinting and ChIP (coupled with the knowledge of ChIP-Seq studies already published in Ref. [183], and comments in the 'discussion' section below) could illuminate the promoter elements that bind or localize to the RGMc promoter. Furthermore, it would be quite useful to look at the RGMc locus in primary hepatocytes in culture over time to determine if there is a "maintenance signal" to keep the RGMc locus open and transcribed in the liver, or if the decrease in RGMc expression in liver cell culture is due to loss of a critical transcription factor as the cells adapt to culture conditions. Clearly, there is a great deal of experimental work to be competed to fully understand the tissue-restricted mechanisms that control the expression of RGMc. These data provide an insight into future experiments discussed in more detail below.

3.5: Discussion

Experiments presented in this chapter delineate the organization of the mouse RGMc gene and define the mechanisms of RGMc gene regulation in skeletal muscle, and provide future insight into experiments to determine the regulatory mechanisms of RGMc expression in the liver. Key findings include the demonstration of alternative RNA splicing between exons 1 and 2, which leads to expression of three distinct RGMc transcripts that vary in the length of the 5' UTR in both striated muscle and liver. Additionally, DNA elements in the RGMc promoter responsible for RGMc gene expression in skeletal muscle were characterized. Comparative analysis of RGMc genes from 10 mammalian species further reveals evidence that both differential RNA processing and the muscle-specific promoter elements have been evolutionarily conserved.

Placing the experimental data in the context of molecular evolution

Many genes undergo alternative RNA splicing [186-189], which can lead to transcripts encoding different protein species [186, 190] or containing distinct regulatory properties, such as differential stability or translatability [191]. The precise mechanisms that control alternative splicing are complex and full understanding of splice-site selection remains incomplete [187]. For RGMc the three mRNAs characterized here vary in the length of the 5' UTR, and appear to be expressed at relatively equivalent levels in muscle and liver. Further studies will be needed to define any distinctive functional properties for each transcript, such as control by cellular iron levels of either RNA stability or translatability (addressed in chapter 4 of this dissertation).

Like several other genes that are expressed in differentiating skeletal myoblasts, RGMc gene transcription is controlled by a combinatorial interplay of muscle-restricted and more broadly expressed transcription factors, many of which are evolutionarily ancient (Fig. 3.22), including members of the myogenic basic helix-loop-helix (bHLH) family, which includes Myf-5, MyoD, myogenin, and MRF-4 [175-177], and the MEF2-family, MEF2A - D [16, 138]. MEF2C has been shown to be directly activated by and to cooperate with myogenic bHLH proteins during muscle differentiation *in vitro* [192], and during skeletal muscle development *in vivo* [193]. Our results demonstrate that a set of paired E-boxes (α-element) and a MEF2 site (γ-element) are important for RGMc promoter activity in muscle cells, and suggest that myogenin and MEF2C may be the key transcription factors acting at these sites (Fig. 3.17). We also have identified a third region in the proximal RGMc promoter, termed the β-element, that also is necessary for full transcriptional activity in differentiating muscle cells, but have not yet defined the responsible transcription factors (Fig. 3.17). Leading candidates include members of the Stat and Ets families, although our preliminary studies rule out Stat5b (Fig. 3.16C). Both Stat3 and several Ets factors will need to be tested, as each have been shown to positively regulate a number of genes expressed during muscle differentiation [194-196].

Besides being produced in skeletal muscle, RGMc is also expressed in cardiac muscle and in hepatocytes (Fig. 3.1, and Ref. [1, 27, 30, 31, 33, 41]). To our knowledge there are no other genes that exhibit a pattern of gene expression that is restricted to just striated muscle and the liver, placing RGMc in a unique position to provide insight into multiple tissue-specific regulatory mechanisms, as well as potential control of whole-body iron metabolism. In this regard, preliminary experiments have shown that the

RGMc promoter fragments identified here are active in skeletal muscle, but are not functional in several liver cell lines (Fig. 3.18), even though a putative binding site for the liver-enriched HNF4α transcription factor is found in the proximal RGMc promoter, and ChIP-Seq data suggest that HNF4α is enriched at the human HJV/RGMc locus in two places (one is at the promoter region; personal communications with, and data from M.D. Wilson of the Odom laboratory, published in Ref. [183])). Further studies will be needed to identify the control elements responsible for RGMc transcription in cardiac muscle and hepatocytes, and to define other regulatory mechanisms, including determining whether or not cellular iron levels influence RGMc gene expression (see chapter 4 of this dissertation). For example, transgenic studies with mutations in the three elements defined here could provide insight into any tissue-specific actions of RGMc, including the possible source of soluble RGMc, which has been detected in the extra-cellular fluid of cultured cells, and in blood [34, 67-72], and has suggested to be a critical regulator of hepcidin gene expression [66, 77-79], and thus iron metabolism.

The results presented in this chapter represent the first detailed analysis of promoter function of any member of the RGM family. As both RGMa and RGMb have completely different profiles of gene expression than RGMc, being produced in distinct parts of the central nervous system, and not in muscle or liver [27-30, 32, 33], it is likely that the critical transcriptional control elements will be different. However, comparative genomic analysis of RGMa and RGMb genes indicates the presence of several well conserved E-boxes and MEF2 sites in their putative promoters (Fig. 3.23B), which could be regulated by neuronal specific bHLH proteins such as N-twist [197] in conjunction with MEF2 family members (shown in Fig. 3.23B as a “critical site X” (in *green*), but this could be any transcription factor binding site that confers tissue-specificity to each RGM family

member). Future studies will be needed to define the specific mechanisms of regulation of RGMa and RGMb gene transcription.

RGMc is expressed in several vertebrates ranging from mice to zebrafish [30, 31, 41, 152], but an ortholog has not been found in avian species to date [152]. A bioinformatic analysis of the RGMc locus in zebrafish reveals several E-boxes, MEF2 sites, Ets/Stat elements (Fig. 3.21), additional stretches of DNA that are conserved between mouse and zebrafish (Fig. 3.23A), and that the putative transcription factors that regulate these elements (i.e., bHLH and MADS-box [MEF2] families) arose early in evolution (Fig. 3.22), suggesting that the promoter is evolutionarily ancient in vertebrates in which RGMc is expressed in muscle. In addition, a single RGM gene has been identified based on genomic data [35, 152], with the strongest evidence being present in the sea squirt, *Ciona intestinalis*. Interestingly, the single RGM in *Ciona* appears to be a 4 exon gene with numerous MEF2, E-box, and Stat/Ets-sites surrounding the putative promoter (Fig. 3.21). For example, a transgenic *Ciona* with the mouse RGMc promoter fully intact versus a promoter with mutations of the three elements defined in this chapter (summarized in Fig. 3.17) could shed light on the evolutionary origins of the RGM family of genes. Whether *Ciona* RGM is expressed in muscle or is specific to the developing embryo or the adult organism is not known at this time. Understanding the regulatory mechanisms of expression of the three RGMs in vertebrates and the single RGM in a model organism like *Ciona* will help provide a full appreciation of the evolutionary history of the entire RGM family and help discern how each member acquired its distinct tissue-specific regulatory modules.

Additional mechanisms of regulation

Gene expression requires a series of tightly regulated steps that often involves a combination of complex modifications at the gene locus. This includes (i) direct binding of transcription factors to the DNA (which has been the primary focus of this chapter), usually in a combinatorial nature, (ii) possible modifications to the chromatin, the complex assembly of DNA, histones, and additional proteins that may bind to the complex, as well as (iii) influences that arise from apparently distal locations; however, the order in which these events occur is still unclear. Furthermore, once the nascent transcript is synthesized, a large number of post-transcriptional processes may act to regulate gene expression (which will be discussed in more detail in chapter 4). Transcriptional regulators must overcome the chromatin barrier to gain access to their sites and affect transcription [198].

The related transcriptional co-activators CBP and p300 (63% identical) [199] have been shown to interact with many of the same transcription factors, but gene knockout studies demonstrate that they regulate different events in development [200]. For example, both CBP and p300 have been found to interact with the sequence-specific transcription factors MEF2 and MyoD [201] in muscle, and promote transcription of the genes with which they are localized by altering the chromatin structure through intrinsic or recruited histone acetyltransferase (HAT) activity. In muscle, p300 is more enriched than CBP [201] and has been shown to activate MyoD through acetylation [202].

The well-described p38-MAPK pathway is used in numerous genes expressed during muscle development, but there are differing views as to the exact order or recruitment. Puri and colleagues suggested that blocking p38 repressed the recruitment of the SWI/SNF complex, but did not affect the chromatin binding of muscle-regulatory factors

and HAT's to those promoters [203, 204]. In contrast, Tapscott and colleagues demonstrate that the SWI/SNF complex (including ATPase Brg1) can be found at the locus of the same genes, and that in the absence of functional SWI/SNF enzymes, the same muscle-regulatory proteins did not bind [205]. I would postulate that this assembly is a possible regulatory mechanism of RGMc regulation as co-activators such as p300 and chromatin-remodeling enzymes such as the SWI/SNF complex are recruited to the myogenin promoter, another gene expressed early in muscle differentiation with similar kinetics as RGMc (Fig. 3.5B and D). CBP and p300 have both been shown to interact with MEF2, HNF4α, and members of the Stat and Ets families of transcription factors, adding another possible regulatory mechanism for RGMc. For example, recruitment of co-activator complexes is also essential in the regulation of genes expressed in the liver, and CBP has been shown to acetylate HNF4α at several lysine residues that are absolutely required for the nuclear retention of the transcription factor [206]. Furthermore, subsequent DNA-binding and transcriptional activity of HNF4α requires the CBP protein [207], however p300 also increases its activity [208] . In contrast, HNF4α has also been associated with HDACs and subsequent gene repression [208], but this association would require HNF4α activation for proper nuclear localization. I would hypothesize that if MEF2 and MyoD are the factors required for RGMc transcription in muscle (as the co-expression experiments in Fig. 3.16 would imply), and HNF4α is required in the liver (as ChIP-Seq [183] and knockout array data [184] would suggest), then the transcriptional co-activators p300 and/or CBP should also be present at the locus in both tissues. Furthermore, if these co-activators are present, then additional chromatin-remodeling molecules, e.g., the SWI/SNF complex, are likely to play an important role in RGMc regulation. As evidenced from above, a great deal of experimental work is needed to elucidate the signaling pathways and downstream affects that regulate the activation of RGMc expression.

In contrast to co-activators, co-repressors have been found to associate with muscle-specific transcription factors. For example, class II histone deacetylases (HDACs), such as HDAC-4 and -5, which are enriched in the brain, heart, and skeletal muscle [209], have been found to bind to MEF2, repressing MEF2-dependent gene activation [210]. For example, methylation of core histones H3K9 and H3K27 by the histone methyltransferases (HMTs) Suv39h1 and E(z) (member of the polycomb group using the Ezh2 recognition module), respectively, are associated with gene repression (reviewed in [11]). Furthermore, five "highly-enriched CpG-islands" within a 40 kb stretch of the mouse RGMc locus (Table 3.2) can be found computationally, creating another possible mechanism to keep RGMc off in tissues that do not express the gene. I would hypothesize that the enriched CpG islands around the RGMc locus may be heavily methylated in tissues that do not express RGMc, and unmethylated in striated muscle and the liver. While the overexpression of these HMTs has been shown to inhibit muscle differentiation [211, 212], their roles *in vivo* remains uncertain.

The overlying theme throughout this dissertation is that of evolution being able to provide insight into the understanding of gene families and subsequent regulation of gene expression and function. The critical assumption is that regions of the genome that have been well-conserved across sufficient evolutionary distance are more likely to be functionally important. While there is a counter example where ultra-conserved regions (defined as regions that show 100% identity over >200 bps between human, mouse and rat genomes [213]) were deleted in transgenic mice with little to no appreciable phenotype [214], there are numerous examples where conservation over a sufficient evolutionary distance provides critical insight into gene function. It is here that I would like to introduce the concept of 'Promoter Synteny,' also called co-linear alignment of 'conserved non-coding elements' (CNEs) [215].

The concept of promoter synteny

In chapter 2, I discussed the molecular evolution of the RGM family using not only DNA sequence data, but location and structure of the genomic loci across multiple species to infer evolutionary relatedness. This method assumes that the genomes are still similar enough that it is possible to align the majority of orthologous sequence at the DNA level yet distant enough that a great deal of variation has had the opportunity to accumulate [216]. By analogy, we can exploit the concept that evolution uses a modular gene organization [15-17] for a series of conserved genomic islands rafting in a sea of genomic change in order to identify features protected from variation by natural selection [36], which have a higher probability of being functionally important. As noted above, all three RGM family members in mouse contain a putative MEF2 site (Fig. 3.23, *red with green outline*) based on sequence alone (see figure legend 3.21 for sequence). The primary challenge at this time is that only the promoter of RGMc has been characterized, and although a MEF2 site may be found in RGMa and RGMb, the sites are ~1 kb from the predicted transcription start site; the functional relevance of the MEF2 sites in RGMa and RGMb are unknown, as neither of their repective promoters have been characterized to date. Understanding the full spectrum of gene family evolution, and expanding this concept to full genome evolution requires comparing modern genomes with ancestral genomes, which thus necessitates the reconstruction of those ancestral genes and genomes [36, 215], and even ancestral proteins to understand function [86, 217, 218].

Since the RGM family of genes have a non-overlapping pattern of expression, the concept that a common transcription factor binding site would at first seem counter-intuitive. However, MEF2 has been shown to interact with a large number of transcription factors, many of which are tissue-specific. For example, hypothetical transcription factors binding sites in the promoters of RGM family members that help to confer tissue-restricted expression of the individual RGM family members are shown in

figure 3.23 (*yellow with a green outline*). Theoretically, this model supports the concept of 'Modular Gene Organization' [15-17] in which the ancestral RGM contains a MEF2 site that is bound and regulated by an evolutionarily ancient MADS-box transcription factor (see Fig. 3.22 and Ref. [219]). Following gene duplication, the individual RGM family members may have used an additional binding site (*yellow/green*) in addition to the MEF2 site to confer the tissue-restricted pattern of expression seen in modern vertebrates that express the three RGM family members. If this hypothesis is correct, it would suggest that the evolutionarily ancient MEF2 trancription factor and TFBS are essential for all RGM family members, from human to zebrafish to sea squirt *Ciona*, and may shed light on the ancestral function of the gene family. This would be particularly strong evidence as the tissue-specific transcriptional regulation of many genes has diverged significantly between human and mouse [220], although it appears that this is highly dependent on the individual genes and gene families being analyzed. While some initial efforts have been made to characterize the natural selection process in promoters [14, 18, 221], clearly, in order to understand the molecular evolution of the regulation of gene families, like the RGM family, the promoters of all the family members must be characterized fully. Generating hypotheses about other family members from well-characterized promoter studies should provide a foundation for more complex studies and expedite the process. Data presented in this dissertation should provide a framework for a full-understanding of the evolution of the RGM family and its mechanisms of regulation.

3.6: Acknowledgements

We thank Lisa Wilson and David Kuninger for assistance with preliminary experiments, and other members of our laboratory for helpful comments during the development of the

work presented in this chapter. The studies reported here have been supported in part by National Institutes of Health grants T32 HL007781 and F30 HL095327 (to C. J. S.), and by R01 DK042748-21 (to P. R.).

Table 3.1: ***Primers used for RT-PCR***

	Gene	Location	DNA Sequence	Product (bp)
Total RNA	RGMc	Exon 3	5' GCACGGTCGAGCCCCGGGCT	282
		Exon 4a	5' GAACCATCTTCAAAGGCTGCAGGAAG	
	myogenin	Exon 1	5' GGGGACCCCTGAGCATTGTCC	612
		Exon 3	5' TGGACATCAGGACAGCCCCAC	
	MEF2C	Exon 5	5' CCCAGTGTCCAGCCATAACAG	344
		Exon 8	5' CAGGTGGGATAAGAACGCGG	
	MyoD	Exon 1	5' TACAGTGGCGACTCAGATGC	312
		Exon 3	5' CTGGGTTCCCTGTTCTGTGT	
	Myosin heavy chain	Exons 37-38	5' TCAGAAACTGGAGACACGGATCAGA	351
		Exon 40	5' AGAGGTGAAGTCACGGGTCTTTGCC	
	Muscle creatine kinase	Exon 7	5' CATGTGGAACGAGCACCTGG	333
		Exon 8	5' TACTTCTGCGCGGGGATCAT	
	S17	Exon 2	5' ATCCCCAGCAAGAAGCTTCGGAACA	302
		Exon 5	5' TATGGCATAACAGATTAAACAGCTC	
Nuclear RNA	RGMc	Exon 1a	5' GGCTGGAGCAGACCAACAGAATAG	212
		Intron 1	5' CAAGAGGAAAAGTGAAGACTGGGG	
	myogenin	Exon 1	5' GGGGACCCCTGAGCATTGTCC	408
		Intron 1	5' CCAAGGGCCCTGCTTTGCACC	
	S17	Exon 2	5' ATCCCCAGCAAGAAGCTTCGGAACA	439
		Intron 2	5' GCCGTCACCAGCCCTCCTCCG	
TSS Mapping	Exon 2 Rev	+1289 : +1308	5' CCCAGATGATGAGCCTCCTA	
	I	-34 : -15	5' CCAACCATATACTCTCCCTC	
	II	-20 : -1	5' TCCCTCCCCCCTCCCCCCAC	
	III	-5 : +18	5' CCCACACCAAACCTCCTCTG	
	IV	+7 : +26	5' CCTCCTCTGGCTCTCTGACC	
	V	+20 : +39	5' TCTGACCTGAGTGAGACTGC	
	VI	+35 : +54	5' ACTGCAGCCATTCCGGGGCA	
5' RACE	Exon 2 Rev	+1289 : +1308	5' CCCAGATGATGAGCCTCCTA	
	Exon 4b Rev	+2961 : +2980	5' TTCAAAGGCTGCAGGAAGAT	
	Exon 1b Rev	+137 : +157	5' TTCCCGAGGTCAAGCATTACT	
	Exon 1c Rev	+110 : +130	5' CTGGGTTCTCGAGCCATAGTT	
	5' RACE Poly(dT)	--	5' GACTCGAGTCGACATCGA(dT)17	
	5' RACE Adaptor	--	5' GACTCGAGTCGACATCG	

Table 3.2: ***Islands of unusually high CG-composition at the RGMc locus***

CpG-enriched site	Start Sequence	End Sequence	Length
I	-10489	-10219	271
II	-4665	-4567	99
III	-578	-513	66
IV	+1470	+1522	53
V (intron 2 & exon 3)	+1859	+2281	423

Sequence numbering relative to RGMc transcription start site in the mouse.

"Islands of unusually high CG composition" is defined by the (expected/observed ratio) via 'CpG Plot' (Ref. [151], from the European Bioinformatics Institute) of a 40 kb locus surrounding RGMc in the mouse.

(This page was intentionally left blank)

Figure 3.1: Patterns of RGMc expression in the embryo and adult. *in situ* hybridization of RGMc mRNA from embryos of mouse, **A** from Ref. [31] (Kuninger, 2004), and zebrafish, **B** acquired via Ref. [41] (Sprague, 2006), at times of development indicated (e = embryonic day; hpf = hours post-fertilization). **C.** RGMc mRNA is expressed in striated muscle and in the liver from an adult (3 mo.) male C57-Bl6 mouse. Results of RT-PCR experiments for RGMc and S17 mRNAs using RNA from mouse tissues (Sk, skeletal muscle (gastrocnemius); Li, liver; K, kidney; Ht, heart; Lu, lung; St, stomach; Br, brain).

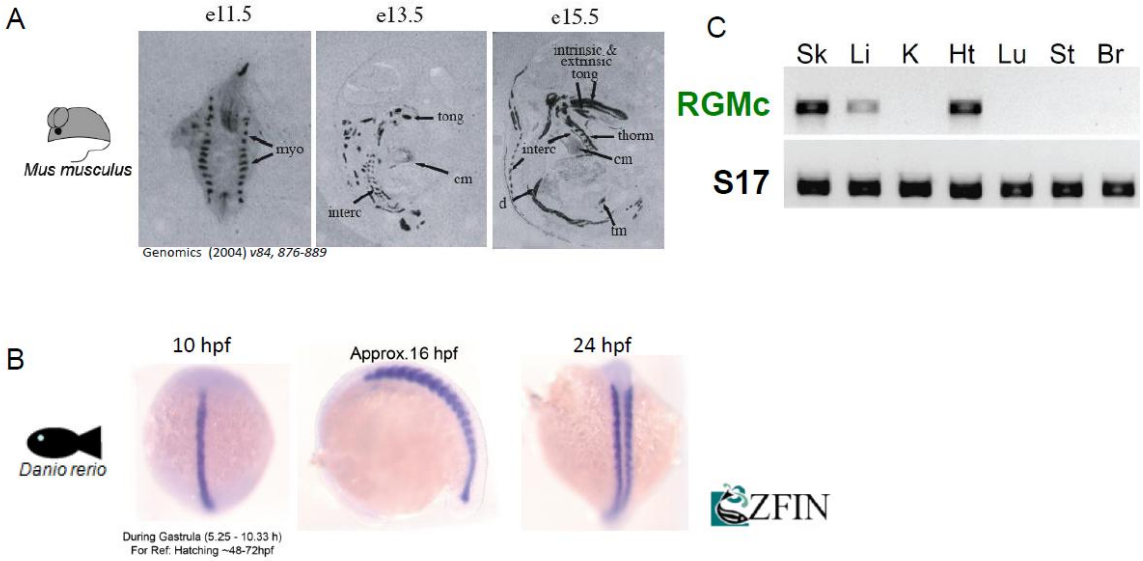

A
e11.5
e13.5
e15.5
Mus musculus
myo
tong
cm
interc
intrinsic &
extrinsic
tong
thorm
cm
d
tm
Genomics (2004) v84, 876-889
C
Sk Li K Ht Lu St Br
RGMc
S17
B
10 hpf
Approx.16 hpf
24 hpf
Danio rerio
During Gastrula (5.25 - 10.33 h)
For Ref: Hatching ~48-72hpf
ZFIN

Figure 3.2: Establishing mouse RGMc gene structure. **A**. Mapping the 5' end of the mouse RGMc gene by 5' RACE using mouse skeletal muscle RNA. The number of clones is graphed on the *y-axis* above the corresponding location of the 5' residue on the x-*axis*. The putative transcription start site is denoted as +1 (*arrowhead*), with exon 1 in *upper case* letters. A potential TATA box is labeled, and primers II - IV used in **(C)** are indicated below the sequence. **B**. Mapping the 5' end of the mouse RGMc gene by RT-PCR with cDNA from mouse skeletal muscle RNA and overlapping PCR primers located in different parts of RGMc exon 1, as seen on the gene map to the left (see Table 3.1 for DNA sequences of primers). Exons 1 and 2 are depicted as boxes, with the 5' UTR in *gray* and the protein coding region in *black*, and introns and flanking DNA as horizontal lines. Results are seen to the right, and molecular weight markers are indicated (see Fig. 3.3A for results with heart and liver RNA). In addition to mapping the 5' end of exon 1, the results also show that alternative RNA splicing occurs between exons 1 and 2. **C**. DNA sequence of the junction between intron 1 and exon 2 of the mouse RGMc gene. Exon 2 is in *upper case* letters; the locations of alternative RNA splicing are noted by *chevrons*, with the –AG splice-acceptor residues *underlined*. **D**. Organization of the mouse RGMc gene and mRNAs. The gene contains 4 exons (*boxes*) and three introns (*thin lines*). The transcription start site is denoted as a *bent arrow*, and the polyadenylation site as a *vertical arrow*. The three RGMc mRNAs are diagramed below, and result from use of alternative splice acceptor sites at the 5' end of exon 2.

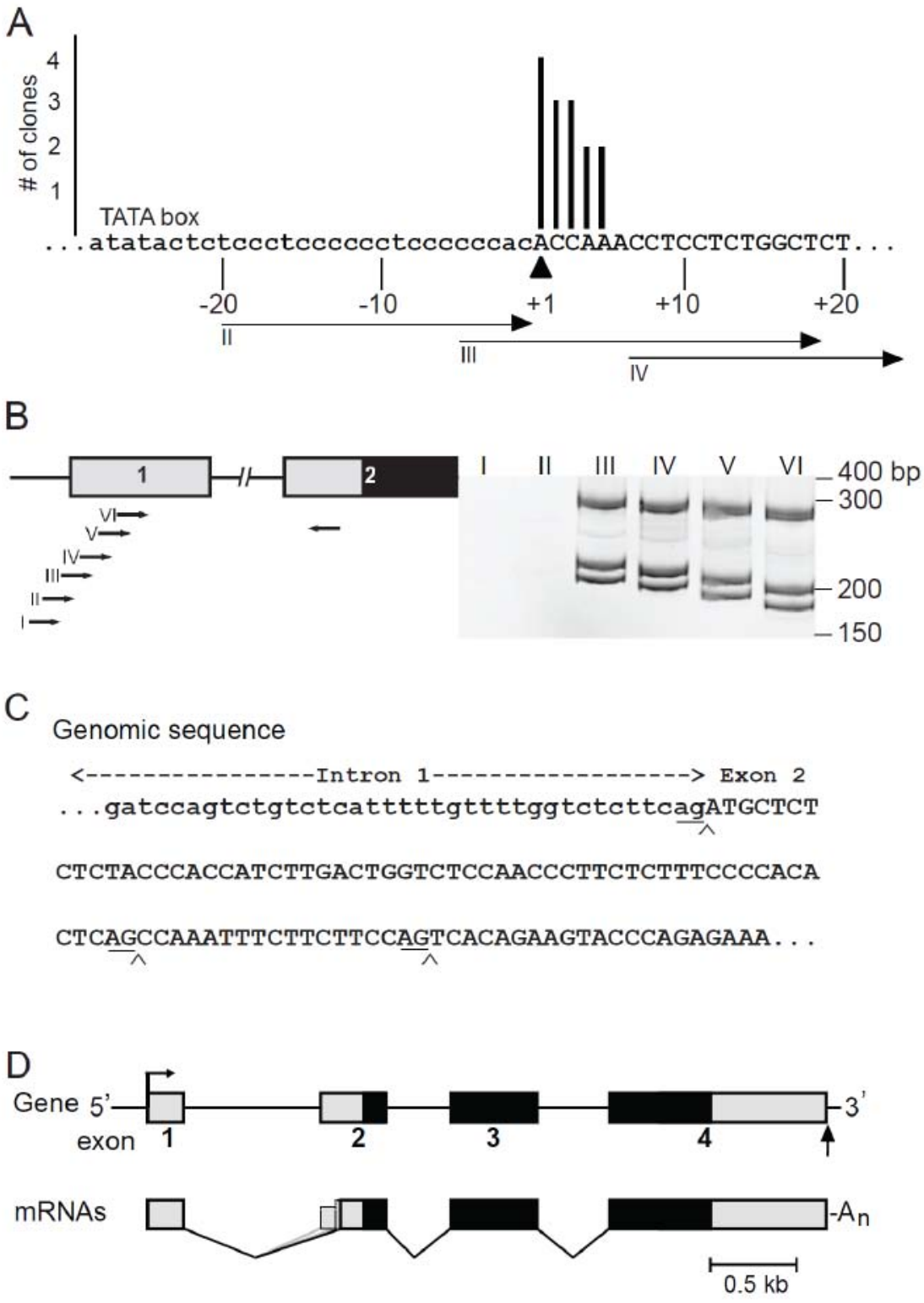
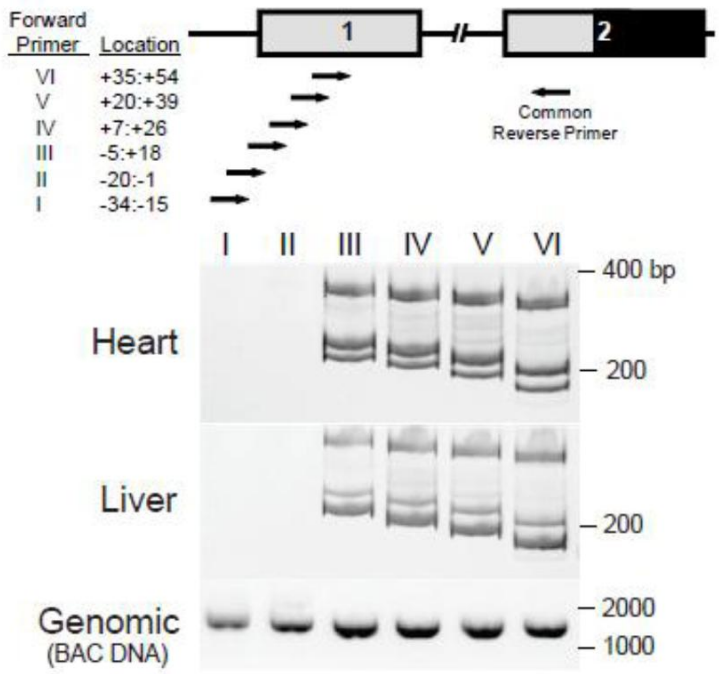

A
# of clones
4
3
2
1
TATA box
...atatactctccctcccccctcccccacACCAAACCTCCTCTGGCTCT...
-20
-10
+1
+10
+20
II
III
IV
B
1
2
VI
V
IV
III
II
I
I
II
III
IV
V
VI
400 bp
300
200
150
C
Genomic sequence
<----------Intron 1----------> Exon 2
...gatccagtctgtctcatttttgttttggtctcttcagATGCTCT
CTCTACCCACCATCTTGACTGGTCTCCAACCCTTCTCTTTCCCCACA
CTCAGCCAAATTTCTTCTTCCAGTCACAGAAGTACCCAGAGAAA...
D
Gene 5'
exon
1
2
3
4
3'
mRNAs
-An
0.5 kb

Figure 3.3. **Alternative RNA splicing involving RGMc exon 2 occurs in heart and liver, and the DNA sequences are conserved among mammalian species**. **A**. Top - Map of RGMc exons 1 and 2, showing locations of primers for RT-PCR experiments below. Exons are depicted as boxes, with the 5' UTR in *gray*, and protein coding sequence in *black*. Bottom - Results of RT-PCR experiments with heart and liver RNA; molecular weight markers are to the right. Below is shown the results of a test of each primer pair with genomic DNA. **B**. Alignment of genomic sequences from 10 mammalian species in the region 5' to RGMc exon 2. Intron 1 is in *lower case* and exon 2 in *upper case*, with mouse exon 2 highlighted as follows: the splice acceptor site that creates 174-nt exon 2 is in *black*, the additional 18-nt that creates the 192-nt variant is in *dark gray*, and the additional 77-nt found in the 251-nt variant is in *light gray*. Also see GenBank accession number AI196626, which shows an EST containing the 192-nt RGMc exon 2 in mouse, and DA762328 and DA764726, which contain similar data for human RGMc.

A



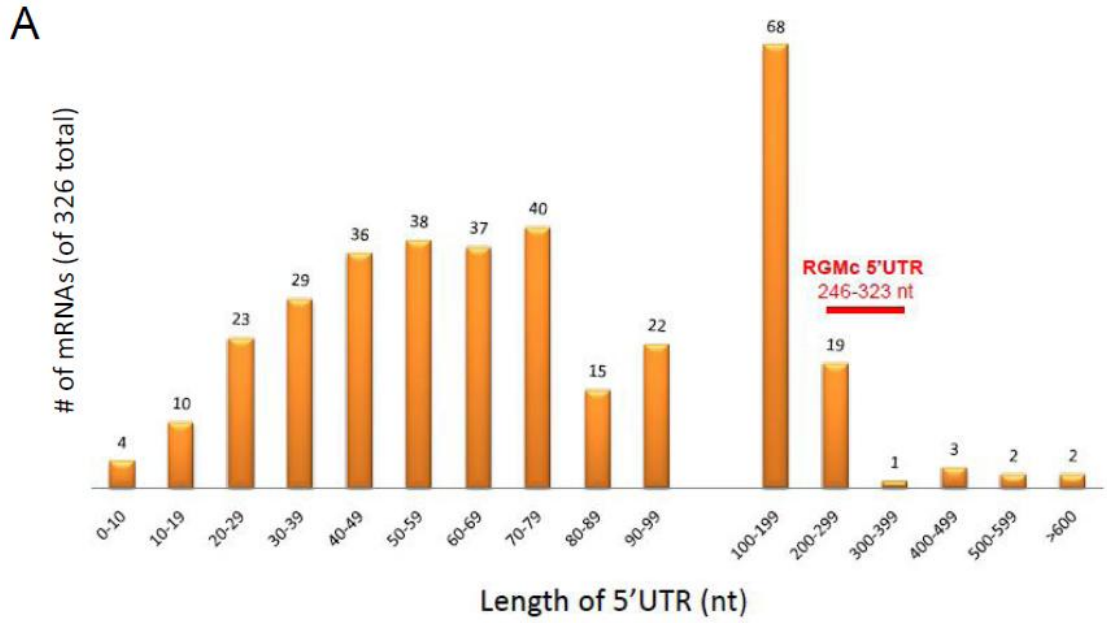
B Alternative RNA splicing in RGMc exon 2

```
                <------------------Intron 1-------------------->    251-nt exon 2 variant
Mouse   tcttcccaagatcc-agtctgtctcattt-ttgttttggtctcttcag ATGCTCTCTCTACCCACCATCT
Rat     tctttccaagatcc-agtctgcttacttt-ttgtcttcatctcctcag at----tctctacccaccatct
Rabbit  tctctccaaaaccc--gactg-ctcatcc-tttccttgatctccccag at----tctcttcacagcatcc
Human   tctccccaaattcc-agtctg-ttcatcc-ttttcttgatctccccag at----tcactccacattatcc
Dog     tttccccaaactcc-cgaccg-ctcatccttttttcctgatcttcccag at----tctcttcatataatcc
Cow     ttcccctaaactcc-agttag-ctcattc-ttttcttgatcttcccag at----tctcttcacatcattc
Armadillo tctccccaaactt---gtctt-cttatgc-tttacttgatctccccaa at----tctcttcacat-accc
Elephant tctccccaaactct-ggtcct-ctcatcc-ttttcttgatcttctctg at----tctcttcacataactc
Tenrec  tctccccagactccaggacct-ctcacct-ttgtcttgagctcctcag at----tcgcttcacg-gaccc
Opossum tttcctttcattcc-tatct--cccatca-tttctttgacctccctag at----cccctataacctttct

                                                    192-nt exon 2 variant
Mouse   TGACTGGTCTCCAACCCTTCTCTTTC-CCCACACTCAG CCAAATTTC---TTCTTCCA----G TCACAG
Rat     tgactaatcttcaacccttctctttc-cccacacgcag ccaaatttc---ttcttccg----a tcgcag
Rabbit  tgaccaatcctcaactcttctctttc-tccatgcccag ccaaatttctctttttttca----g tcactt
Human   ttaccaatcttcaattcttctctctc-tccatgtccag ccaaatttc---ttttttca----g tcactt
Dog     tgaatgatcttcgactcttctctctc-cccatgcccag ccaagtttc---ttttttca----g tccttt
Cow     tgaacaatcttcaactcttctct-tc-cccatgcccag ccacatttc---ttttttca----g tccctt
Armadillo tgaccaatcttcaactcttctctctcgtctatgcccag ccaaatttc---ttttttca----g tccctt
Elephant tgaccaatcttcaactcttctttctcccctatgcccag ccaaatttc---ttttttca----g tccctt
Tenrec  cgcctgacc----acccttctttctccgctatgcccag ccagatttc---cttcctca----g cccctt
Opossum cactcagactttggctcctcttaccc--caacttccaa ccagatttc---cctttaaacaagg tccctt

                                    Exon 2 (174-nt variant)
Mouse   A-AGTA---CCCAGAGAAATTCACTAGGTAGGA-GGCTCATCATCTGGGAAGAACCGGTG-CCTGGGG...
Rat     a-agtagtactcggcgaaattcactaggtagga-ggctcctcatctgggaagaaccggtg-cctgggg...
Rabbit  acaggacagccggtcaaaattcactaggtaggagggctcatcagctgggaagaatcgggg--ctgggg...
Human   acagggcttccggtcaaaattcactaggtagga-gggtcatcagctgggaagaaccggcg-cctggga...
Dog     acaggaggtccggtcaaaattcactaggtagga-gggtcgtcagctggaaaaagccggag-cctgggg...
Cow     acaggacgtccggtcaaaattcactaggtagga-gggtcatcagctggaaagaaccggaa-cctgggg...
Armadillo acaggacgaccggtcaaaattcactaggtagga-aggtcatcagctggaaagagccggag-cctgggg...
Elephant acaggacgtctggtcaaaattcactaggtagga-gggtcatcagctggaaagaaccggag-cctgggg...
Tenrec  acaggacctccggcccaaatccactaggtagga-gagtcatcagctggaaggaaccggag-cctgggg...
Opossum ccagggaatccagccag-------cagccggaa-g---------------agaactggggaccagggg...
```

Figure 3.4. **Size distribution of UTRs in eukaryotes**. **A**. Summary of data from Ref. [154] (Kozak, 1987) of the 5' untranslated region (UTR) length from 346 mRNAs for which the transcription start site has been mapped. For genes with multiple transcripts, only the longest UTR was scored. The range of UTR size is listed on the *x*-axis and number of mRNAs found within that range on the *y*-axis. **B**. Graphical summary of genomics data from Ref. [155] (Pesole, 2001) representing the range of 5' and 3' UTRs from several eukaryotic genomes. Range of UTR sizes is indicated on a logarithmic scale, *y-axis*, with the average length indicated with a *triangle*. For both **A.** and **B.**, the range of 5'UTR length of RGMc mRNAs (from figs. 3.2 and 3.3) are depicted in *red*.


A
# of mRNAs (of 326 total)
4
10
23
29
36
38
37
40
15
22
68
19
1
3
2
2
RGMc 5'UTR
246-323 nt
0-10
10-19
20-29
30-39
40-49
50-59
60-69
70-79
80-89
90-99
100-199
200-299
300-399
400-499
500-599
>600
Length of 5'UTR (nt)

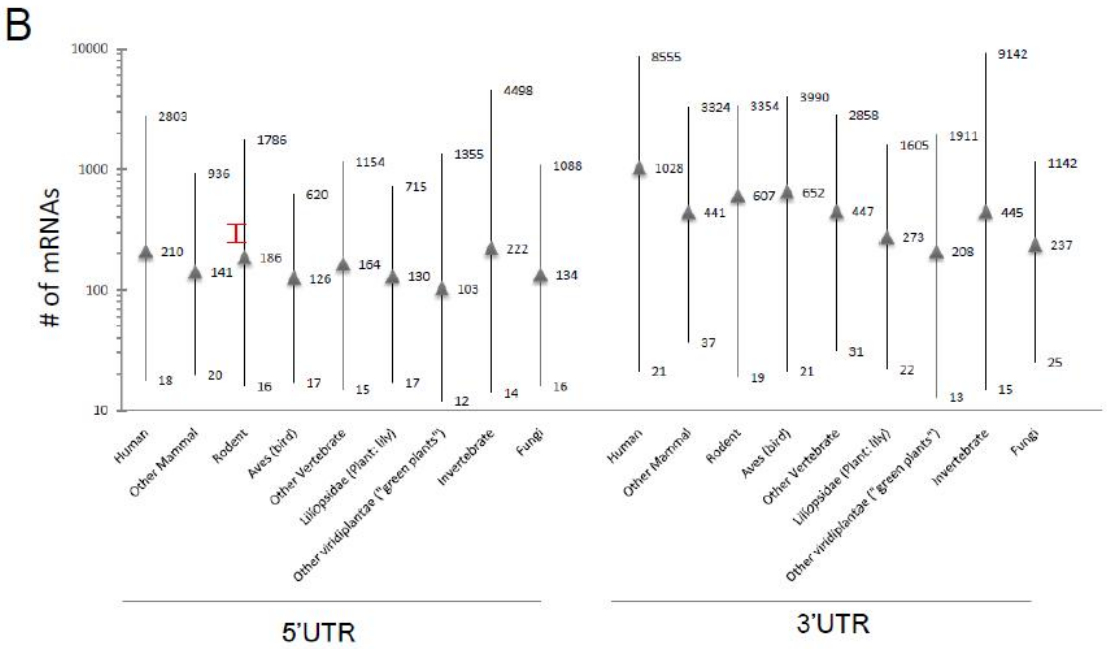
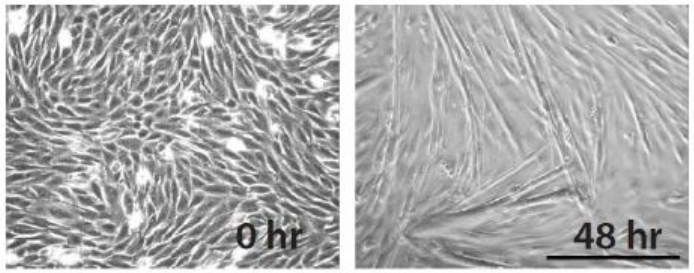
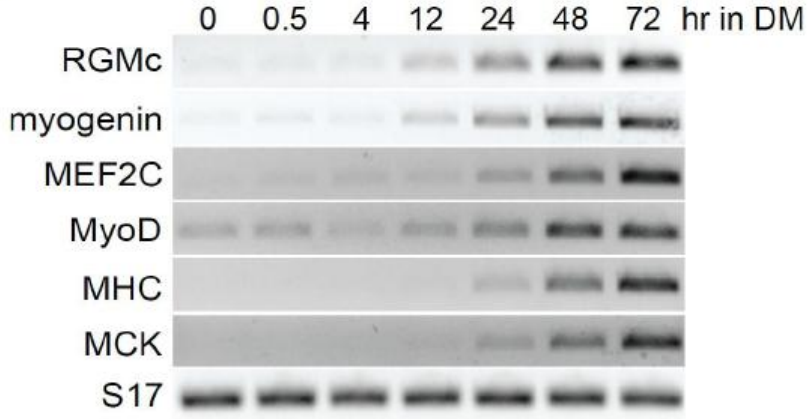

B
# of mRNAs
10000
1000
100
10
2803
210
18
936
141
20
1786
186
16
620
126
17
1154
164
15
715
130
17
1355
103
12
4498
222
14
1088
134
16
8555
1028
21
3824
441
37
3354
607
19
3990
652
21
2858
447
31
1605
273
22
1911
208
13
9142
445
15
1142
237
25
Human
Other Mammal
Rodent
Aves (bird)
Other Vertebrate
Liliopsidae (Plant: lily)
Other viridiplantae (*green plants*)
Invertebrate
Fungi
5'UTR
3'UTR

Figure 3.5. RGMc gene transcription is induced during skeletal muscle differentiation. **A**. Myotube formation occurs during C2 myoblast differentiation, as illustrated by phase contrast images (200X magnification) at confluent cell density (0 hr) and after incubation in DM for 48 hr. Scale bar is 250 μm. **B**. Time course of gene expression for RGMc, myogenin, MEF2C, MyoD, myosin heavy chain (MHC), muscle creatine kinase (MCK), and S17 during C2 myoblast differentiation measured by RT-PCR. **C**. Diagram of the method for obtaining levels of nascent transcription. Briefly, cells were harvested at time points indicated and the nuclear fraction was isolated via centrifugation. Nascent RNA transcripts were subsequently isolated from the nuclei, followed by semi-quantitative RT-PCR using intron-exon primer pairs. See *Experimental Procedures* for detailed information. **D**. Time course of RGMc, myogenin, and S17 gene transcription during C2 myoblast differentiation, as measured by accumulation of nascent nuclear transcripts. Gene maps are to the *right*, and show approximate locations of intron-exon primer pairs (see Table 3.1 for DNA sequences of primers). Exons appear as black boxes (exon sizes are not to scale).

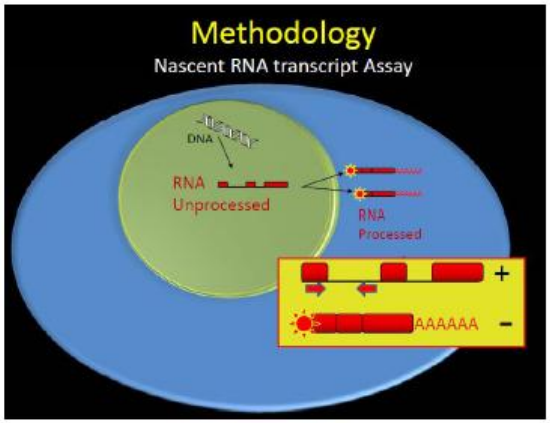
A



B mRNA levels



C



D

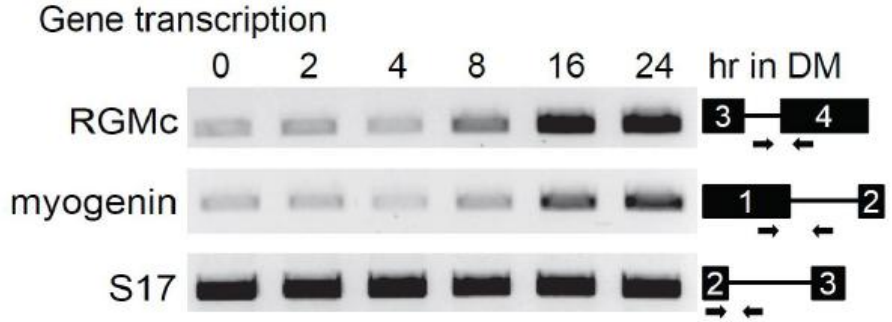
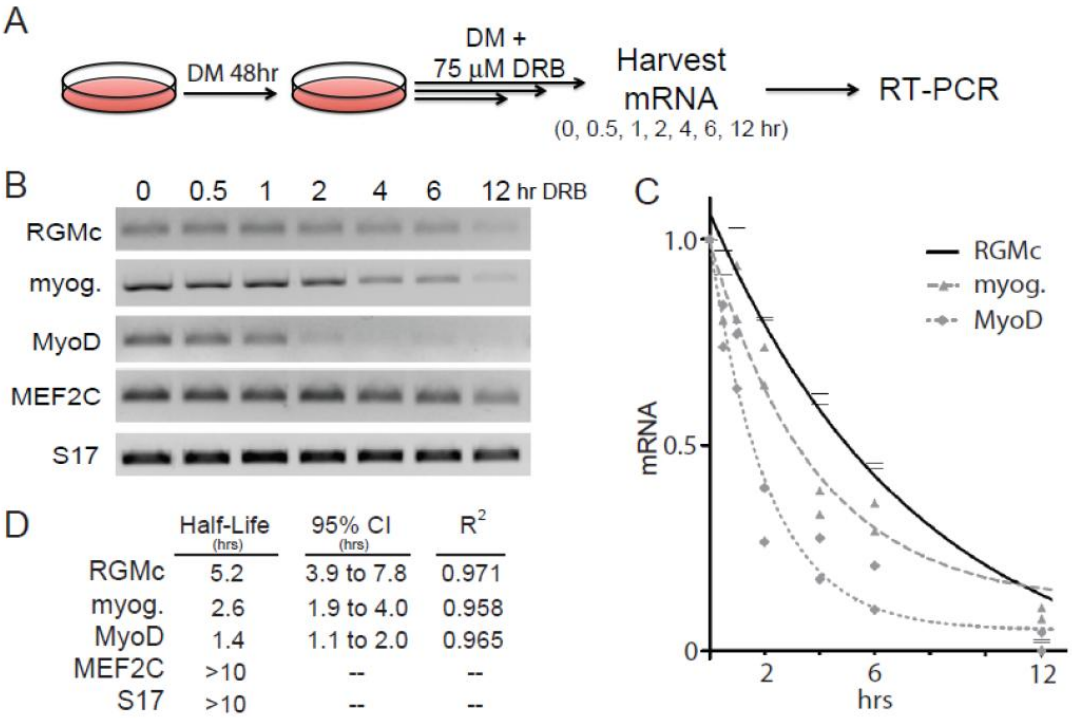


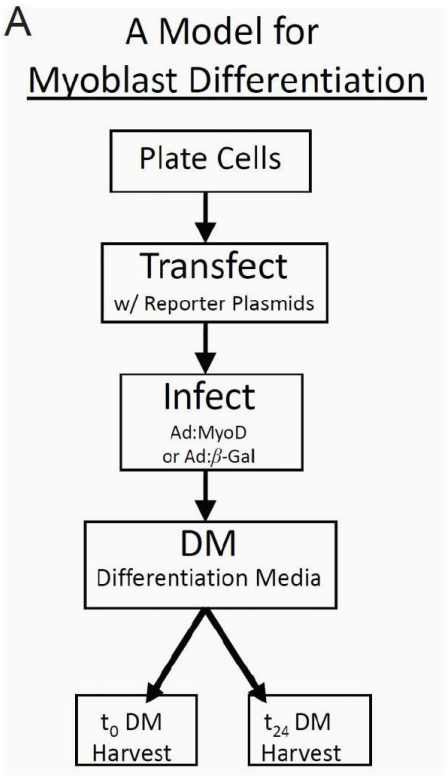
Figure 3.6. mRNA half-life of genes in differentiating C2 myoblasts. Measurement of the half-life of mRNA of RGMc, myogenin, MyoD, MEF2C, and S17 genes in C2 myoblasts. **A**. Experimental protocol. C2 cells grown to confluent cell density were incubated in DM for 48 hr followed by treatment with the transcription elongation inhibitor DRB (75 μM) in DM. Cells were harvested for mRNA at times indicated and subjected to RT-PCR to obtain the mRNA decay rate. **B.** Representative mRNA decay results of RT-PCR of the approximate midpoint of the linear-phase of amplification from two independent experiments. **C.** Graph showing the densitometry results of two independent biological experiments in relation to a mechanistic model of mRNA decay using non-linear regression fit to a one-phase decay equation: $Y = Y_0 \bullet e^{-k \bullet X}$. Values at time 0 hrs set to 1.0, and all other results normalized to this value. **D.** Table showing estimated half-life derived from non-linear curves in **C**, as well as the 95% confidence interval and R-squared value for data on each curve. The half-lives for MEF2C and S17 are listed as greater than 10 hours, as the t½ is beyond the limits of the experiment (due to cell death from transcriptional inhibition at 24 hours). See '*Experimental Procedures*' for details.

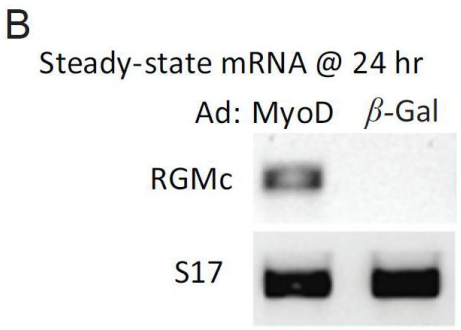


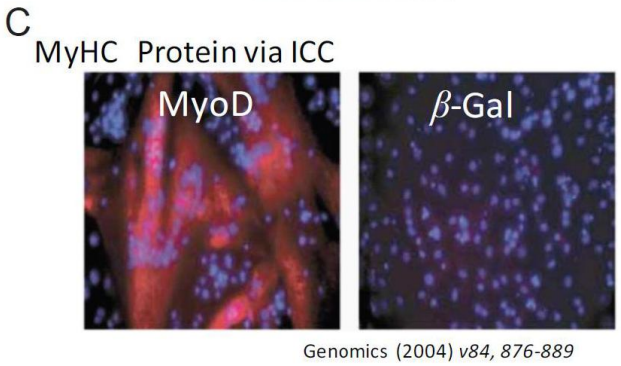
D

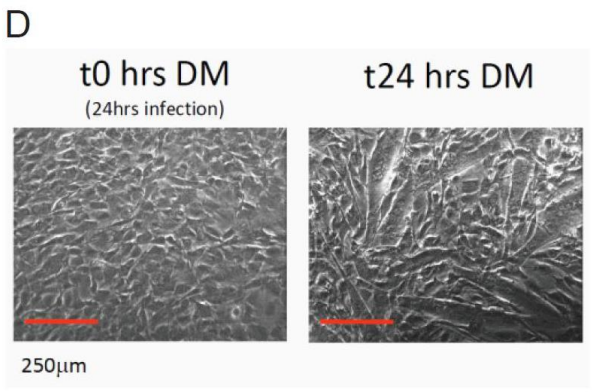
	Half-Life (hrs)	95% CI (hrs)	R^2
RGMc	5.2	3.9 to 7.8	0.971
myog.	2.6	1.9 to 4.0	0.958
MyoD	1.4	1.1 to 2.0	0.965
MEF2C	>10	--	--
S17	>10	--	--

Figure 3.7. A robust model for muscle differentiation: expression of RGMc in cells infected with adenovirus expressing the muscle-specific transcription factor MyoD. Mouse mesenchymal C3H:10T½ cells are infected with adenovirus expressing MyoD, or β–gal control, and induced to differentiate into the myoblast lineage under low serum conditions. Cells that proceed toward the muscle-lineage express RGMc whereas those cells that become fibroblasts do not. **A**. Experimental protocol for muscle differentiation. Details may be found in '*experimental methods*.' **B**. RT-PCR from total mRNA isolated from 10T½ cells infected with Ad:MyoD or Ad:β–gal, and amplified using RGMc or ribosomal S17 primers. **C.** Immunocytochemistry of 10T½ cells infected with Ad:MyoD or Ad:β–gal and stained for myosin heavy chain (*red*) and Hoechst 33258 nuclear dye (*blue*). Data from Ref. [31] (Kuninger, 2004). Myotube formation and in Ad:MyoD-10T½ cells with RGMc expression occuring during myoblast differentiation, as illustrated by **D.** phase contrast microscopy (200X magnification) at confluent cell density (0 hr) and after incubation in DM for 24 hr (Scale bar is 250 μm) and **E.** via western blot of myogenin, MEF2, MyoD, and α-tubulin from whole cell extracts as described in [136].


A
A Model for
Myoblast Differentiation
Plate Cells
Transfect
w/ Reporter Plasmids
Infect
Ad:MyoD
or Ad:β-Gal
DM
Differentiation Media
t0 DM
Harvest
t24 DM
Harvest


B
Steady-state mRNA @ 24 hr
Ad: MyoD
β-Gal
RGMc
S17


C
MyHC Protein via ICC
MyoD
β-Gal
Genomics (2004) v84, 876-889


D
t0 hrs DM
(24hrs infection)
t24 hrs DM
250μm

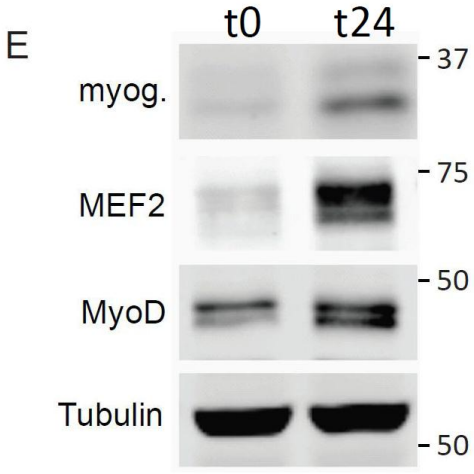
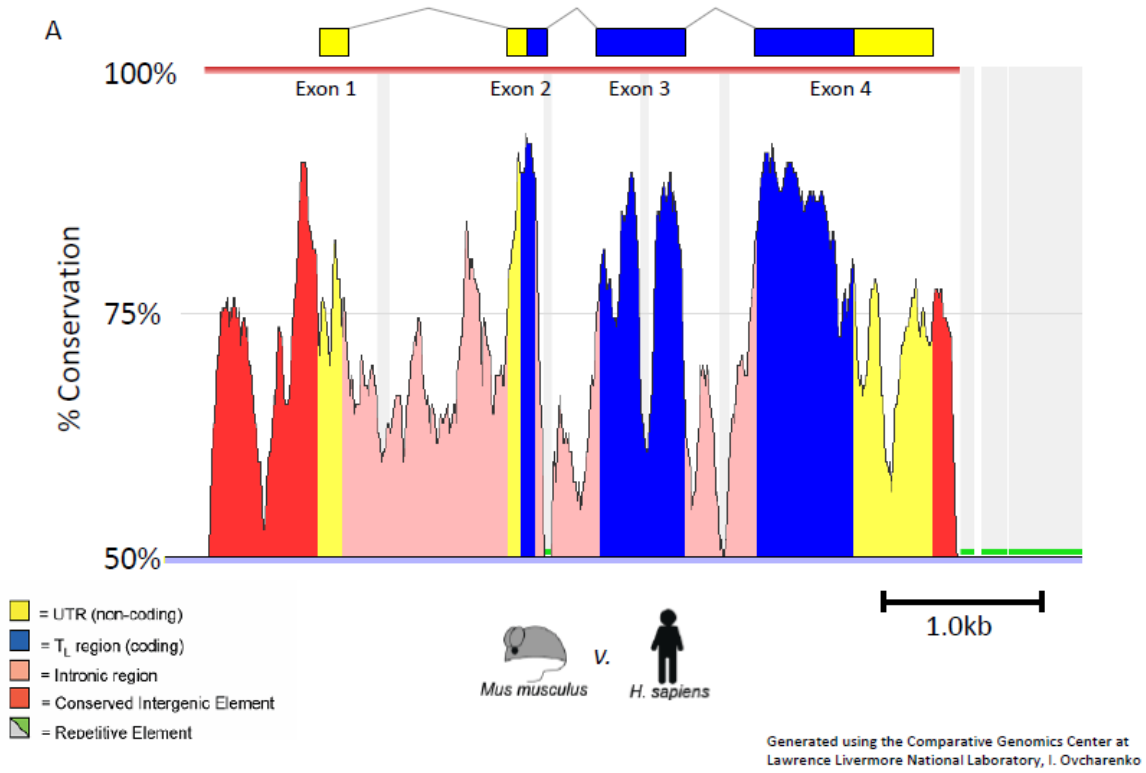

E
t0
t24
myog.
MEF2
MyoD
Tubulin
37
75
50
50

Figure 3.8. Genomic conservation of RGMc/HJV locus between mouse and human. Comparing the RGMc genomic sequence of mouse, *Mus musculus*, to human, *Homo sapiens,* (from Feb. 2006). **A**. Percent conservation of the RGMc locus derived from the Comparative Genomics Center at Lawrence Livermore National Laboratory. Graph shows percent conservation over a sliding window of 100 bps, *y*-axis, over the RGMc locus, *x*-axis. Four exons for RGMc are shown as boxes above the graph, with the following color scheme: non-coding/untranslated sequence in *yellow,* translated regions shaded *blue,* intronic regions in *salmon*, with conserved (>50%) intergenic regions shown in *red*, and repetitive elements in *green* and shaded *gray*. Scale bar is approximately 1 kb of sequence. Regions below 50% conservation are not shown. **B.** Percent conservation upstream of the transcription start site (TSS) broken into regions of ~5 kb (out to 15 kb upstream of the TSS), ~1 kb, and ~100 base pairs. Below the percentage is the number of nucleotides by which the alignment was calculated as well as the % of gaps (and number of gaps over total sequence) for that region. Alignment created using the EMBOSS Needleman-Wunsch (GLOBAL) alignment algorithm utilizing the EDNAFULL substitution matrix from the European Bioinformatics institute (EBI) with a Gap Penalty of 10.0 and Extension Penalty of 0.5.


A
100%
75%
50%
% Conservation
Exon 1
Exon 2
Exon 3
Exon 4
= UTR (non-coding)
= T_L region (coding)
= Intronic region
= Conserved Intergenic Element
= Repetitive Element
Mus musculus
V.
H. sapiens
1.0kb
Generated using the Comparative Genomics Center at
Lawrence Livermore National Laboratory, I. Ovcharenko

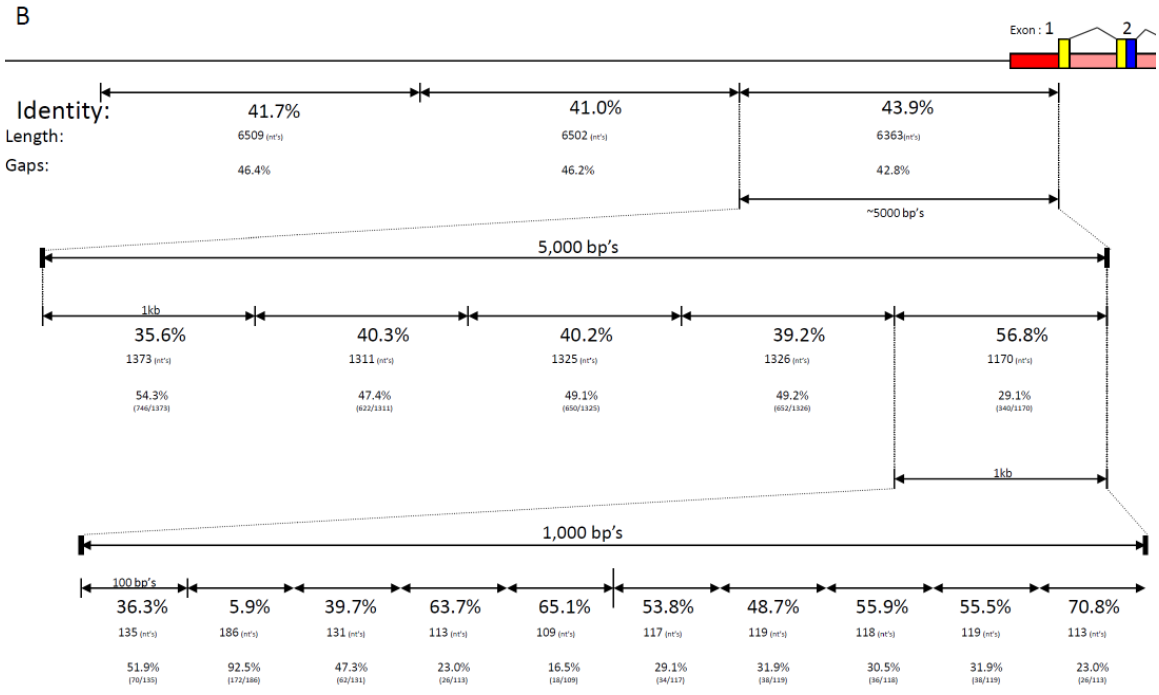

B
Exon : 1
2
Identity:
Length:
Gaps:
41.7%
6509 (nt's)
46.4%
41.0%
6502 (nt's)
46.2%
43.9%
6363 (nt's)
42.8%
~5000 bp's
5,000 bp's
1kb
35.6%
1373 (nt's)
54.3%
(746/1373)
40.3%
1311 (nt's)
47.4%
(622/1311)
40.2%
1325 (nt's)
49.1%
(650/1325)
39.2%
1326 (nt's)
49.2%
(652/1326)
56.8%
1170 (nt's)
29.1%
(340/1170)
1kb
1,000 bp's
100 bp's
36.3%
135 (nt's)
51.9%
(70/135)
5.9%
186 (nt's)
92.5%
(172/186)
39.7%
131 (nt's)
47.3%
(62/131)
63.7%
113 (nt's)
23.0%
(26/113)
65.1%
109 (nt's)
16.5%
(18/109)
53.8%
117 (nt's)
29.1%
(34/117)
48.7%
119 (nt's)
31.9%
(38/119)
55.9%
118 (nt's)
30.5%
(36/118)
55.5%
119 (nt's)
31.9%
(38/119)
70.8%
113 (nt's)
23.0%
(26/113)

Figure 3.9. RGMc promoter activity is induced during muscle differentiation. Results of luciferase assays in differentiating Ad-MyoD-10T½ cells and C2 myoblasts that were transiently transfected with reporter genes containing the minimal thymidine kinase (TK) promoter, the mouse myogenin promoter, or the mouse RGMc promoter (coordinates -4204/+118), and incubated in DM for 0 (*white bars*), or 24 or 48 hr (*gray bars*). The graphs summarize results of ≥3 independent experiments (mean ± S.E.), each performed in duplicate (* - $p < 0.05$, ** - $p < 0.005$ vs. t 0). Myogenin promoter values at t 0 have been set to 100 in each graph (average measurements at t 0 were 7.8×10^4 (Ad-MyoD-10T½ cells) or 7.3×10^3 (C2 cells) relative light units/μg total protein/sec).

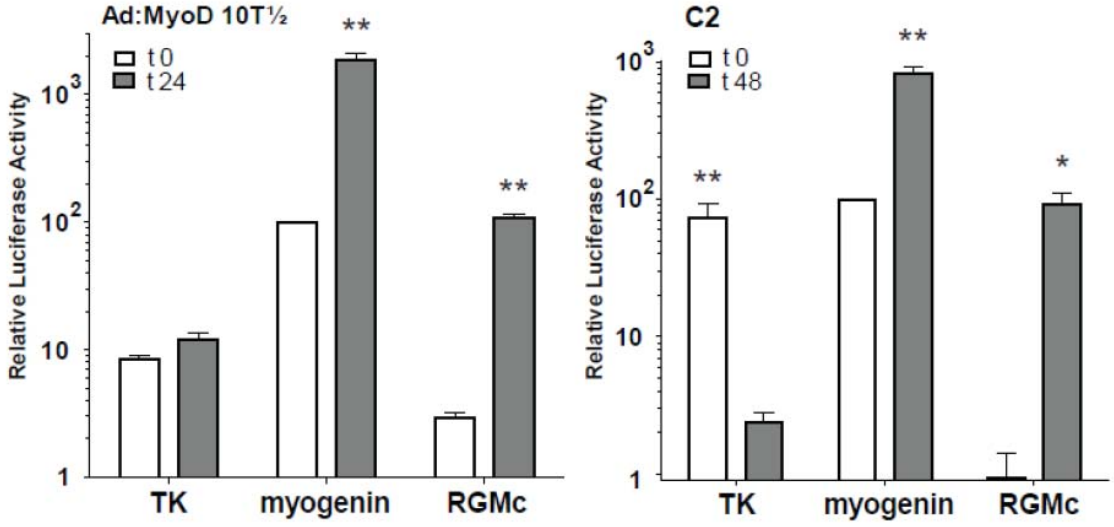

Ad:MyoD 10T½
t 0
t 24
Relative Luciferase Activity
10^3
10^2
10
1
**
**
TK
myogenin
RGMc
C2
t 0
t 48
Relative Luciferase Activity
10^3
10^2
10
1
**
**
*
TK
myogenin
RGMc

Figure 3.10. Mapping regions of the RGMc promoter that activate gene transcription during muscle differentiation. The graphs show results of luciferase assays (mean ± S.E. of 3 - 5 independent experiments each performed in duplicate) in differentiating Ad-MyoD-10T½ cells (*left panel*) and C2 myoblasts (*right panel*) transiently transfected with reporter genes containing a series of 5' truncations of the mouse RGMc promoter. Cells were incubated in DM for 0 (*white bars*), or 24 or 48 hr (*gray bars*) before analysis (* - $p < 0.05$, ** - $p < 0.005$).

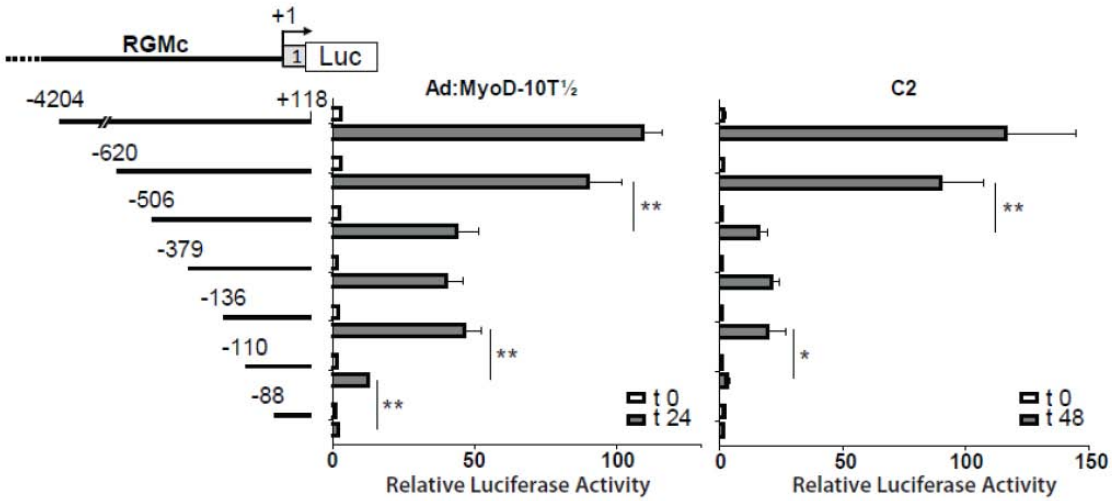
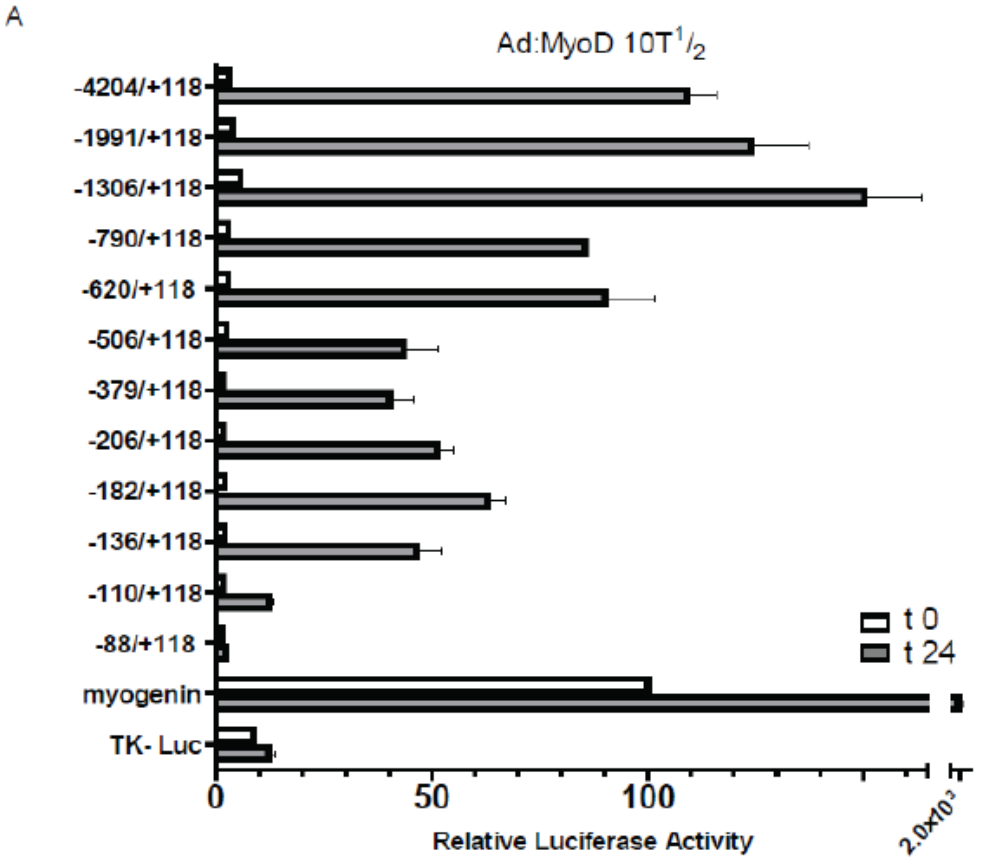

+1
RGMc
1
Luc
-4204
+118
-620
-506
-379
-136
-110
-88
Ad:MyoD-10T½
C2
**
**
**
*
t 0
t 24
t 0
t 48
0
50
100
150
Relative Luciferase Activity
Relative Luciferase Activity

Figure 3.11. **Detailed mapping of mouse RGMc promoter elements in differentiating muscle cells**. The graphs show results of luciferase assays in differentiating Ad-MyoD-10T½ cells (**A**) and C2 myoblasts (**B**) transiently transfected with reporter genes containing different 5' truncations of the mouse RGMc promoter. Cells were incubated in DM for 0 (*white bars*), or 24 or 48 hr (*gray bars*) before analysis (mean ± S.E. of $\geq$ 3 independent experiments, each in duplicate, except the single experiment with -790/+118 construct in 10T½ cells).


A
Ad:MyoD 10T1/2
-4204/+118
-1991/+118
-1306/+118
-790/+118
-620/+118
-506/+118
-379/+118
-206/+118
-182/+118
-136/+118
-110/+118
-88/+118
myogenin
TK- Luc
t 0
t 24
0
50
100
2.0x10³
Relative Luciferase Activity

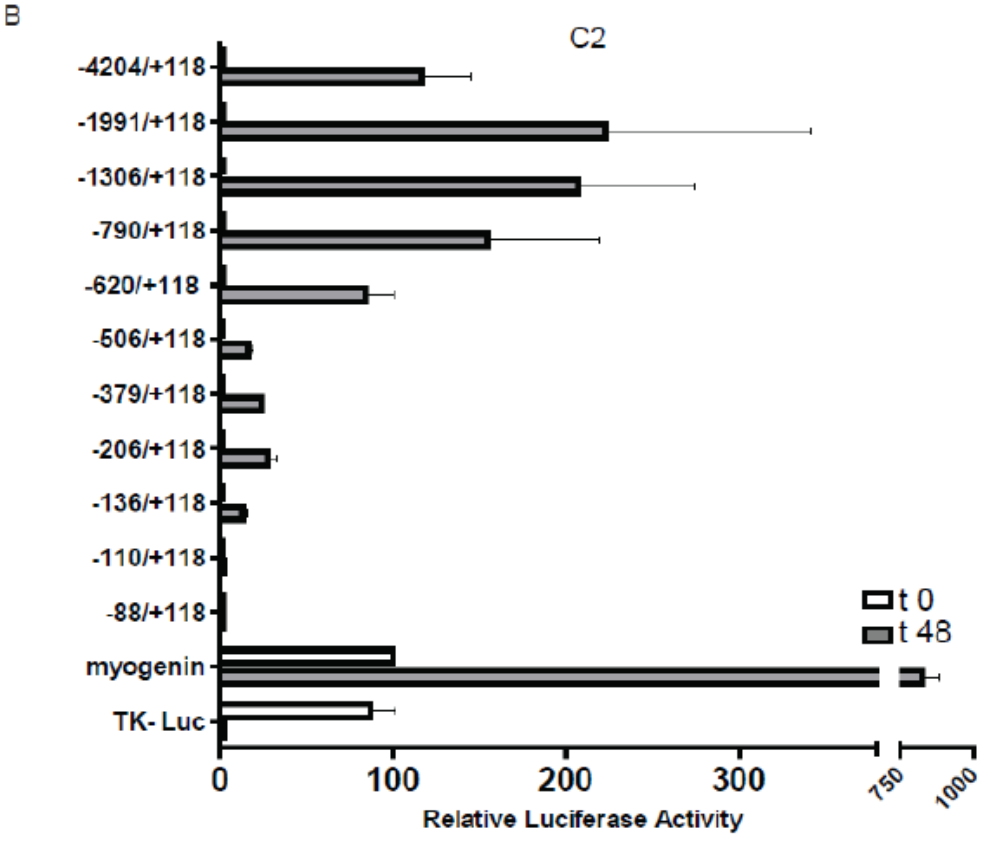

B
C2
-4204/+118
-1991/+118
-1306/+118
-790/+118
-620/+118
-506/+118
-379/+118
-206/+118
-136/+118
-110/+118
-88/+118
myogenin
TK- Luc
t 0
t 48
0
100
200
300
750
1000
Relative Luciferase Activity

Figure 3.12. Analyzing the RGMc gene for potential transcriptional enhancers. **A**. Percent conservation of the RGMc locus between mouse RGMc and human HJV derived from the Comparative Genomics Center at Lawrence Livermore National Laboratory. The graph shows percent conservation as a 'Pip-plot', whereby horizontal length of the line corresponds to the length of the alignment and includes alignment gaps, *x-axis*, while its *y-axis* corresponds to the percent identity for that segment between mouse and human. RGMc exons are shown as boxes above the graph, with the following color scheme: non-coding/untranslated sequence in *yellow,* translated regions shaded *blue,* intronic regions in *salmon*, with conserved (>50%) intergenic regions shown in *red*, and repetitive elements in *green* and shaded *gray*. Scale bar is approximately 4 kb of sequence. Regions below 50% conservation are not shown. **B**. Map of mouse RGMc gene showing regions that were fused downstream of firefly luciferase (Luc) and the RGMc promoter (coordinates -620 to +118) to test for enhancer activity in differentiating Ad-MyoD-10T½ cells. **C**. Graphs depict results of luciferase assays after incubation in DM for 0 (*white bars*) or 24 hr (*gray bars*) (mean ± S.E. of 3 experiments, each performed in duplicate; values for the RGMc promoter at 24 hr were set to 100 (* - $p < 0.01$, ** - $p < 0.001$, vs. t 0).

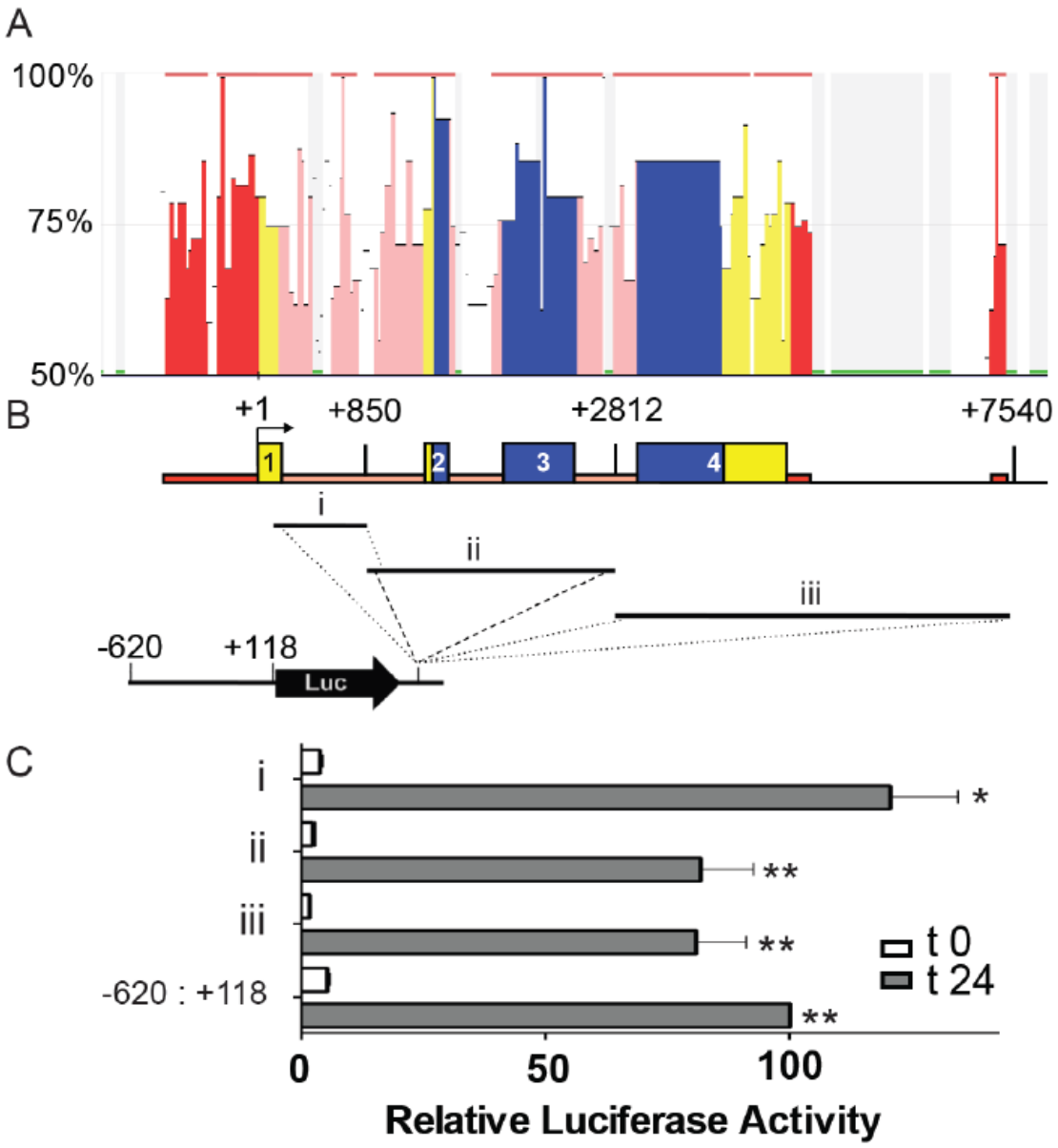

A
100%
75%
50%
B
+1
+850
+2812
+7540
1
2
3
4
i
ii
iii
-620
+118
Luc
C
i
ii
iii
-620 : +118
*
**
**
**
t 0
t 24
0
50
100
Relative Luciferase Activity

Figure 3.13. Characterizing promoter elements that control RGMc gene transcription during muscle differentiation. **A.** Changes via site-directed mutagenesis to the RGMc promoter are grouped as following: the α-site as a set of paired E-boxes, the β-site at a putative stat and/or ets element, γ-site is a MEF2 element. Wild-type (WT) sequence is listed above each region as well as the mutated nucleotides below (Mut) in *bold* and *italics*. **B.** Results are depicted of luciferase assays in differentiating Ad-MyoD-10T½ cells (*left panel*) and C2 myoblasts (*right panel*) transiently transfected with reporter genes containing substitution mutations of the mouse RGMc promoter (illustrated on the maps to the *far left*; details including full primer sequences are in '*Experimental Procedures'*). Cells were incubated in DM for 0 (*white bars*), or 24 or 48 hr (*gray bars*) before analysis. The graphs depict results of 3 - 10 independent experiments (mean ± S.E.), each performed in duplicate (* - $p < 0.05$, ** - $p < 0.001$, vs. t 0).

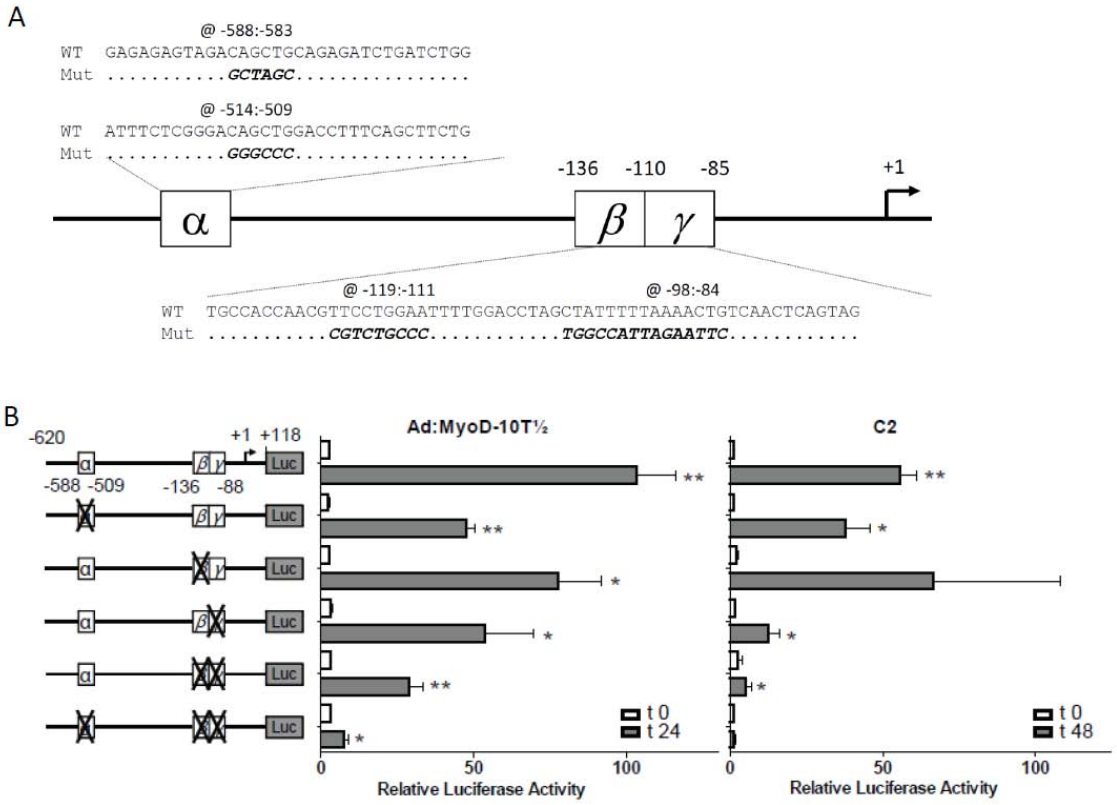
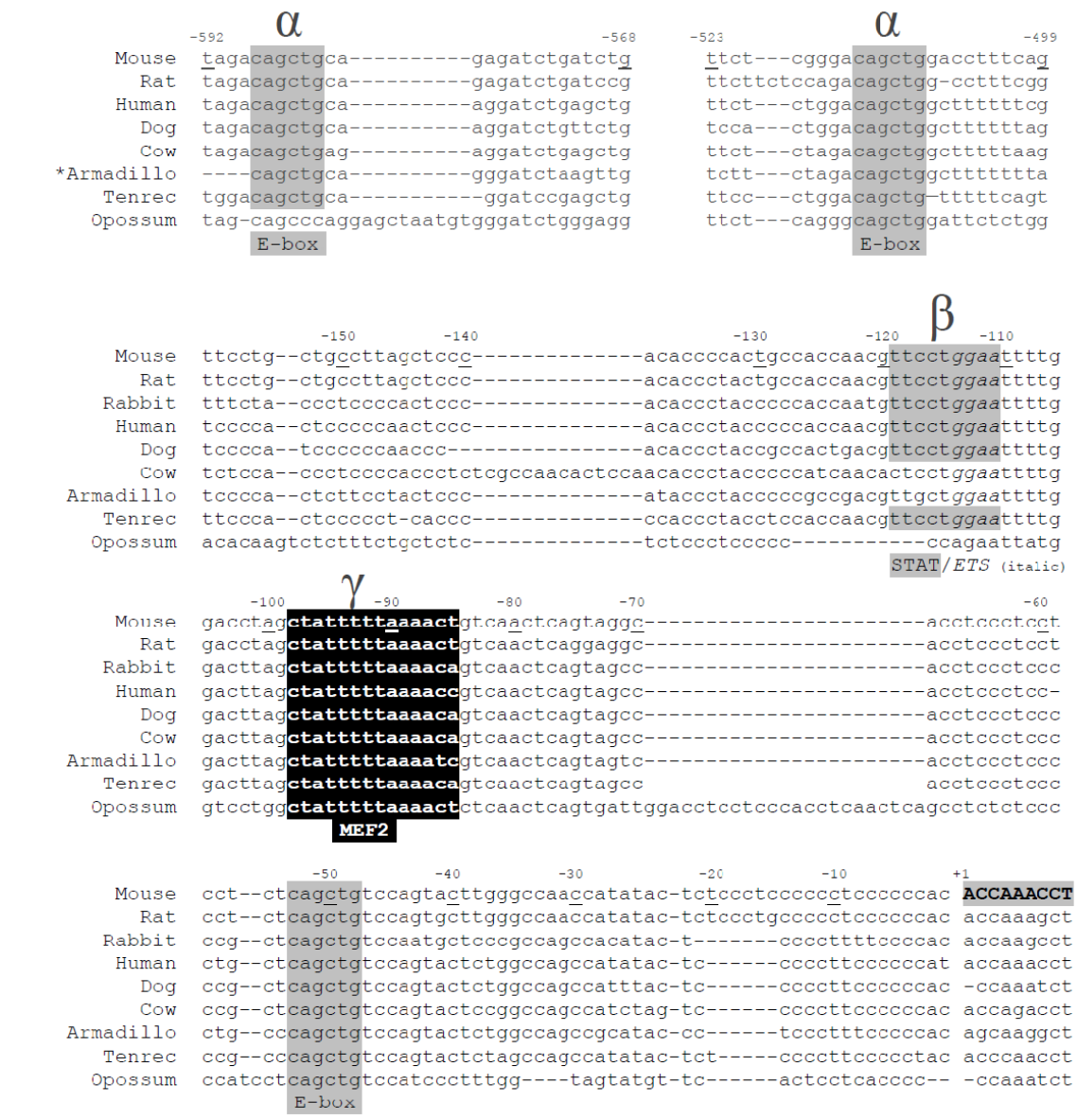

A
@ -588:-583
WT GAGAGAGTAGACAGCTGCAGAGATCTGATCTGG
Mut .....GCTAGC.....
@ -514:-509
WT ATTTCTCGGGACAGCTGGACCTTTCAGCTTCTG
Mut .....GGGCCC.....
-136 -110 -85
+1
α
β
γ
@ -119:-111
@ -98:-84
WT TGCCACCAACGTTCCTGGAATTTTGGACCTAGCTATTTTTAAAACTGTCAACTCAGTAG
Mut .....CGTCTGCCC.....TGGCCATTAGAATTC.....
B
-620
+1 +118
-588 -509
-136 -88
Luc
Ad:MyoD-10T½
C2
t 0
t 24
t 48
0
50
100
Relative Luciferase Activity

Figure 3.14. Comparative mapping of RGMc promoter elements from different species. DNA sequence alignment of part of the proximal RGMc promoter from 8 mammalian species. Highlighted regions include paired E-boxes at -588 to -583 and -514 to -509 (α-site), the β-site from -120 to -110, and the γ site, a putative MEF2 element, from -98 to -85. Another E-box also is indicated. Mouse exon 1 is in *upper case* and *bold* letters.



* Note, large gaps in Armadillo sequence, Rabbit Sequence not available for this region

Figure 3.15. Binding of the γ-element by nuclear protein extracts. Preliminary electrophoretic mobility shift assays (EMSA) to define the protein-DNA interactions at the RGMc promoter. Probe with the asterisk, '*', is shown in part (**D**), with other probes showing a low level of binding (data not shown). **A.** Schematic overview of the main elements (*red*) of the RGMc proximal promoter and relative position of the double-stranded oligos used in the preliminary EMSA assays. Regions that showed a drop in promoter activity (see Fig. 3.10 and 3.13) are labeled as II and III, with approximate locations of three oligos that span the region positioned below. Infrared (IR) phosphoramidite-labels are shown as spheres at the 5' end of the oligo, with the 3' end depicted as a *'closed arrowhead.*' **B.** Purity of the nuclear extracts. Western blots of Akt (Pan Akt; Cell Signaling #9272, Rb Poly) and CREB (Upstate UBI 06-863, Rb Poly) from (25 μg) cytoplasmic and (5 μg) nuclear fractions following nuclear extraction of C2 cells at t0 and t48 hrs in DM, and 10T½ fibroblast cells. See '*experimental procedures'* for details of high salt extraction. **C.** Results of gel-mobility shift assays using infrared (IR)-labeled double-stranded oligonucleotides for MEF2 control element in the mouse muscle creatine kinase (mMCK) promoter. (*Left*) MEF2-IR-oligos with nuclear protein extracts from C2 cells at t0 and t48 hrs DM (*lanes3 and 6*, respectively) and 10T½ fibroblasts (*lane 8*). This distinct doublet is more pronounced in C2 cells that have undergone differentiation (*compare lanes 3 and 6)*, and may be competed off using an unlabeled ('cold') MEF2-oligo using 1x (*lane 4*) or 10x (*lanes 5, 7, and 9*) molar ratio of unlabeled to labeled probe. (*Right*) IR-labeled MEF2 probe with C2 nuclear extract alone (*lane 3*), with 10x unlabeled MEF2 (*lane 4*), and 10x non-specific (ns; Oct1) probe (*lane 5*). **D.** Comparing the MEF2 site in the RGMc gamma element (*lanes 1 and 2*, probe with asterisk '*' in (**A**)) versus mMCK MEF2 site (*lanes 3-7*). The banding pattern of the RGMc probe with a MEF2 site is similar to the MEF2 site found in the mMCK promoter (*compare lanes 2 and 5*). NS = non-specific (Oct1) unlabeled double-stranded oligo probe.

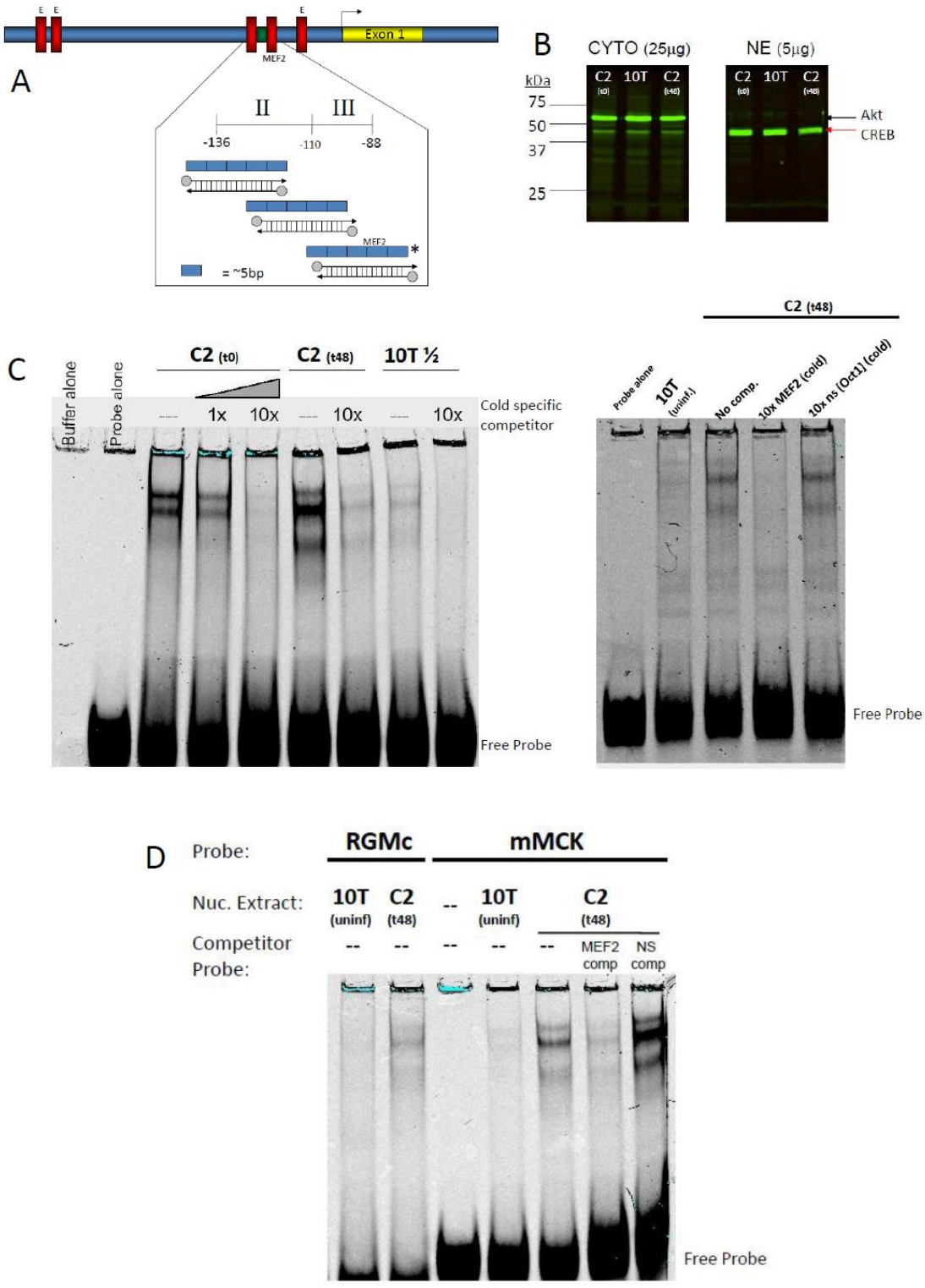

A
E E
E
Exon 1
MEF2
II
III
-136
-110
-88
MEF2
= ~5bp
B
CYTO (25µg)
NE (5µg)
kDa
75
50
37
25
C2 10T C2
Akt
CREB
C
C2 (t0)
C2 (t48)
10T ½
Buffer alone
Probe alone
1x
10x
10x
10x
Cold specific competitor
Free Probe
C2 (t48)
Probe alone
10T (uninf.)
No comp.
10x MEF2 (cold)
10x ns (Oct1) (cold)
Free Probe
D
Probe:
RGMc
mMCK
Nuc. Extract:
10T (uninf)
C2 (t48)
--
10T (uninf)
C2 (t48)
Competitor Probe:
--
--
--
--
--
MEF2 comp
NS comp
Free Probe

Figure 3.16. Stimulation of RGMc promoter activity by myogenin and MEF2C. Effects of myogenin (**B**), stat5b (**C**), or MEF2C (**D**) on the activity of wild-type (wt) or mutant (depicted as an *X*) RGMc promoters in 10T½ fibroblasts. **A**. Conceptual overview of co-expression experiments, adapted from [4] (Lodish, 2000). **B**. Results of luciferase assays after co-transfection of a myogenin or EGFP (vector) expression plasmid with either wt RGMc promoter, the α-mutant, or 4xE-box-Luciferase control plasmid. **C**. Results of luciferase assays after co-transfection of a constitutively active STAT5b or pcDNA3 (vector) expression plasmid with either wt RGMc promoter, the β-mutant, or IGF-I HS-7 luciferase control plasmid demonstrating that the RGMc promoter is not repsonsive to STAT5b. **D**. Results of luciferase assays after co-transfection with an expression plasmid for constitutively active MEF2C (MEF2C-VP16) or pcDNA3 (vector) and either wt RGMc promoter or the γ-mutant. For **B** through **D** the graphs represents results of 3 independent experiments (mean ± S.E.), each performed in duplicate (* - $p < 0.05$, ** - $p < 0.005$).

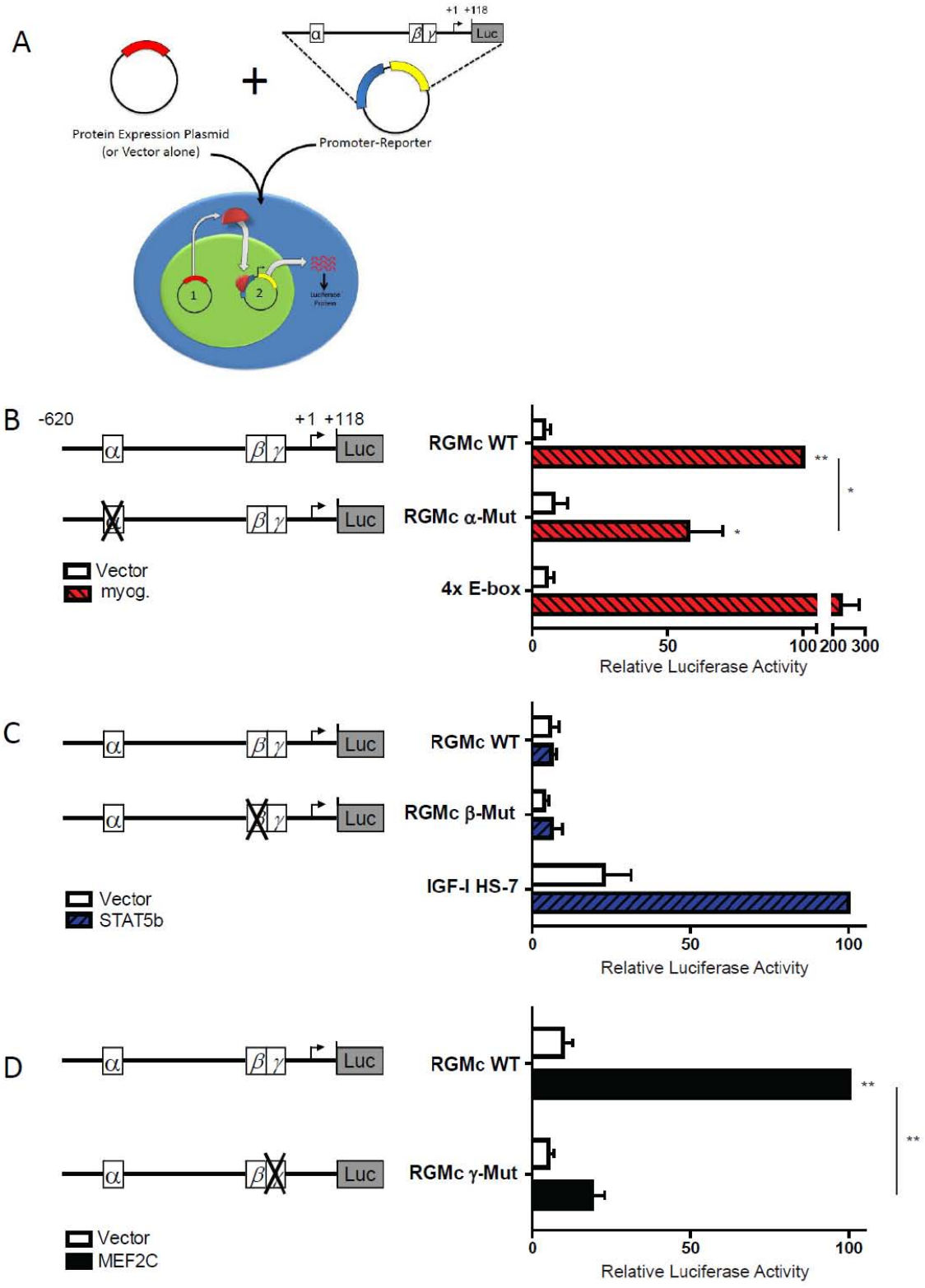
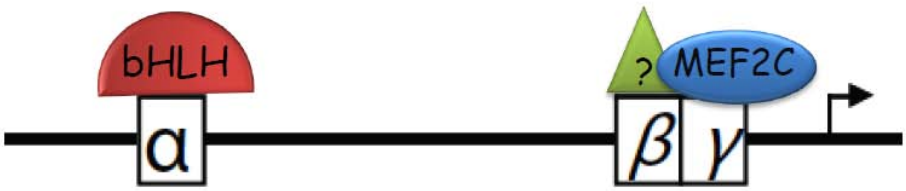

A
+1 +118
α
βγ
Luc
Protein Expression Plasmid
(or Vector alone)
Promoter-Reporter
1
2
Luciferase Protein
B
-620
+1 +118
Vector
myog.
RGMc WT
RGMc α-Mut
4x E-box
0
50
100
200
300
Relative Luciferase Activity
C
Vector
STAT5b
RGMc WT
RGMc β-Mut
IGF-I HS-7
0
50
100
Relative Luciferase Activity
D
Vector
MEF2C
RGMc WT
RGMc γ-Mut
0
50
100
Relative Luciferase Activity

Figure 3.17. **Model of the RGMc promoter in skeletal muscle**. The proximal RGMc promoter used during muscle differentiation contains 3 core DNA regions which include paired E-boxes at -588 to -583 and -514 to -509 (α-site), the β-site from -120 to -110, and the γ site, a MEF2 element, from -98 to -85. Another conserved E-box also is found at -53 to -48 (See fig. 3.14). The transcription start site (TSS) is denoted as a *bent arrow*. The α-sites appear to be regulated by transcription factors of the basic-helix-loop-helix family (bHLH), e.g., myogenin, shown as a *red semi-circle*, while the γ site is likely bound by a member of the MEF2 family, shown as a *blue oval*. The factor(s) that control the β-element are unknown at this time. See text for additional discussion.

The RGMc Promoter in Skeletal Muscle



=

- Another STAT (e.g., STAT3)
- ETS (e.g., PU.1/SPI)
- NFAT

Figure 3.18. **The RGMc proximal promoter is not active in 3 unique liver cell lines**. **A.** Phase-contrast microscopy of normal mouse AML-12 cells (*left*), as well as human hepatoma HepG2 (*center*) and Hep3B (*right*) cells. Scale bar is 250μm. **B.** Results of luciferase assays in AML-12 (*gray*), HepG2 (*light green*), or Hep3B (*dark green, hatched*) cells transiently transfected with reporter genes containing different 5' truncations of the mouse RGMc promoter. Cells were incubated in various growth media (see '*experimental procedures*' for details) for 24 hr before analysis (error bars = range of 2 independent experiments in duplicate for RGMc constructs and n=6 for control plasmids, IGF-IIp3 and TK-luciferase). Values were normalized to protein concentration and TK-luciferase values were set to 100 relative luciferase activity units (average measurements were 8 $x10^3$ (AML-12 cells), 10 x 10^4 (HepG2 cells), and $3x10^4$ (Hep3B cells) light units/μg total protein/sec).

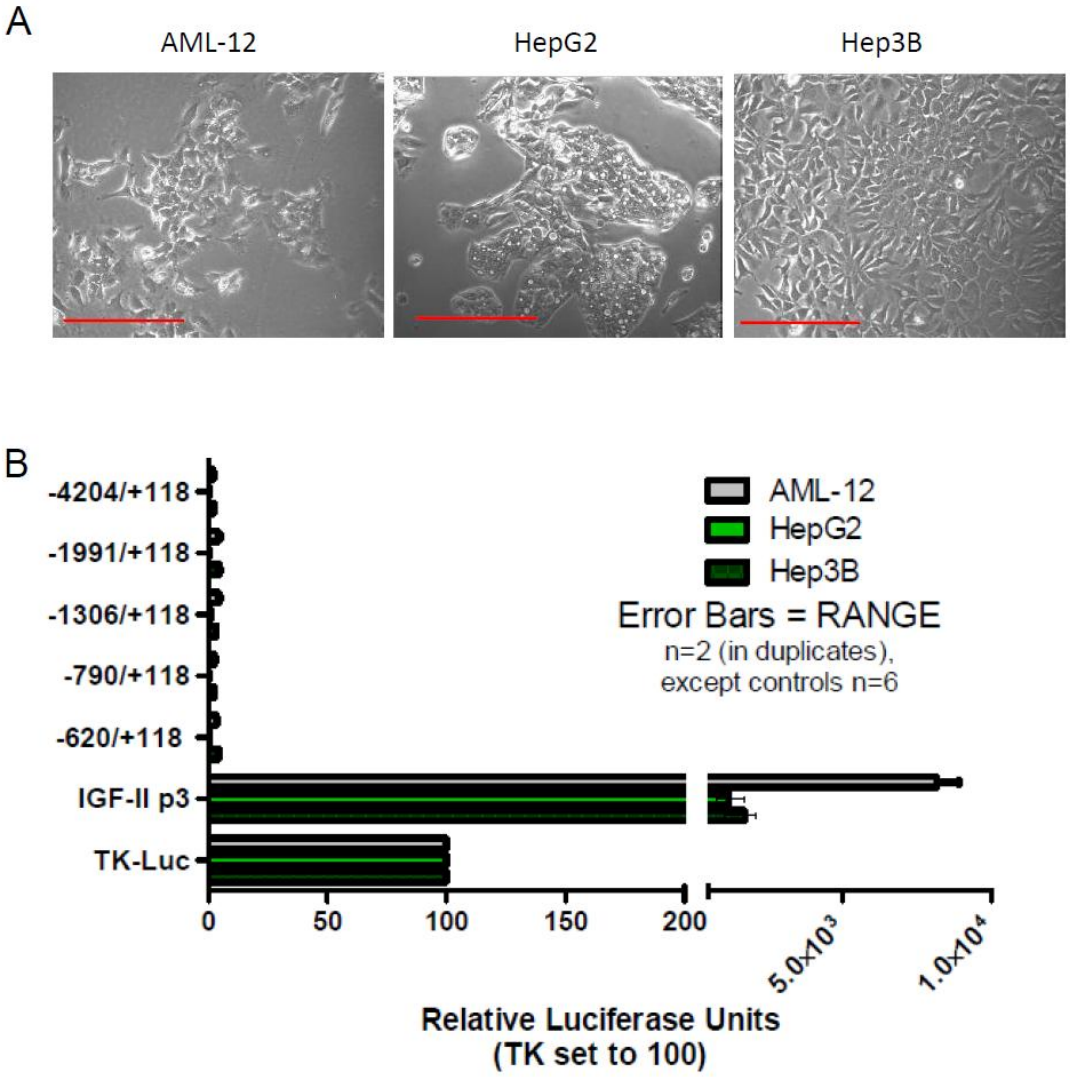
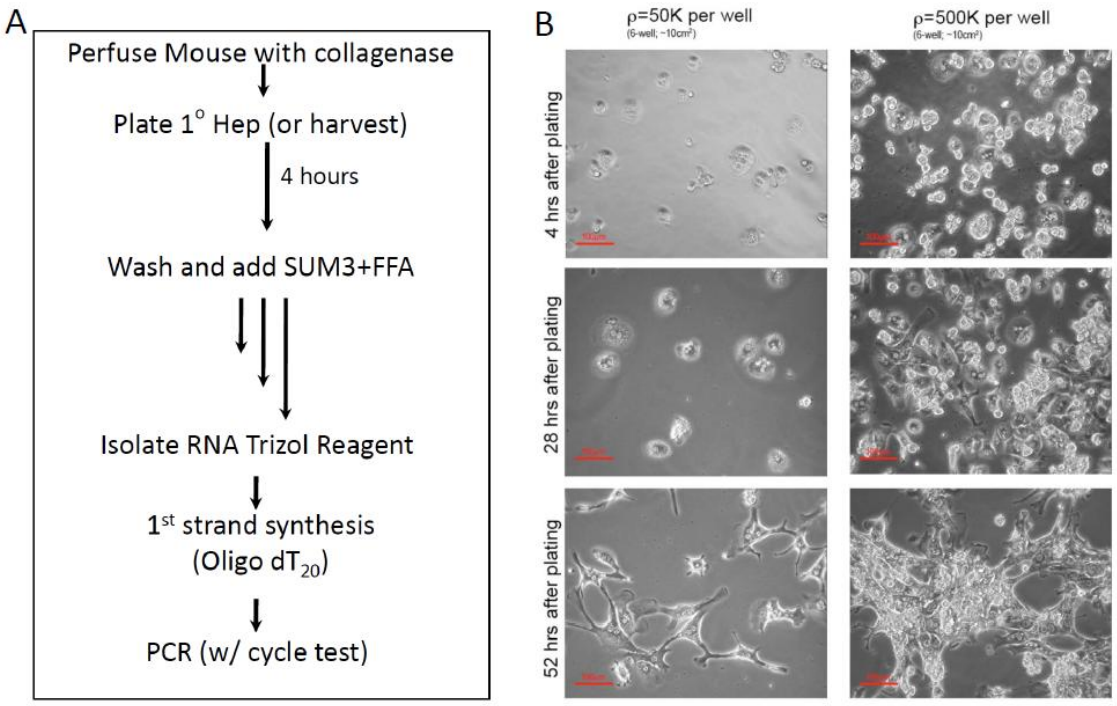

A
AML-12
HepG2
Hep3B
B
-4204/+118
-1991/+118
-1306/+118
-790/+118
-620/+118
IGF-II p3
TK-Luc
AML-12
HepG2
Hep3B
Error Bars = RANGE
n=2 (in duplicates),
except controls n=6
0
50
100
150
200
5.0x10³
1.0x10⁴
Relative Luciferase Units
(TK set to 100)

Figure 3.19. Primary hepatocytes express RGMc mRNA for at least 52 hours in culture. **A.** Experimental overview. SUM3, is a media for primary cultures along with free-fatty acids (FFA). Harvest of primary hepatocytes was performed by A. Duncan of the Grompe Lab, OHSU. All subsequent work was performed by the author of this dissertation. **B.** Morphology of primary hepatocytes at $5x10^4$ and $5x10^5$ cells per well (~$10cm^2$ surface area) over the indicated incubation times. Scale bar, 100 μm. **C.** Results of RT-PCR experiments for RGMc, albumin, and S17 mRNAs using RNA from mouse tissues, isolated primary hepatocytes (pre., pre-plating cells, and house after plating cells are indicated), and rat FTO-2B cells. Note: RGMc and S17 cycles are semi-quantitative, and close to previously established cycle numbers to tissues, whereas the albumin lanes are likely saturated with respect to cycle number and should be taken as an indication of present or absent.


A
Perfuse Mouse with collagenase
Plate 1° Hep (or harvest)
4 hours
Wash and add SUM3+FFA
Isolate RNA Trizol Reagent
1st strand synthesis
(Oligo dT20)
PCR (w/ cycle test)
B
ρ=50K per well
ρ=500K per well
4 hrs after plating
28 hrs after plating
52 hrs after plating

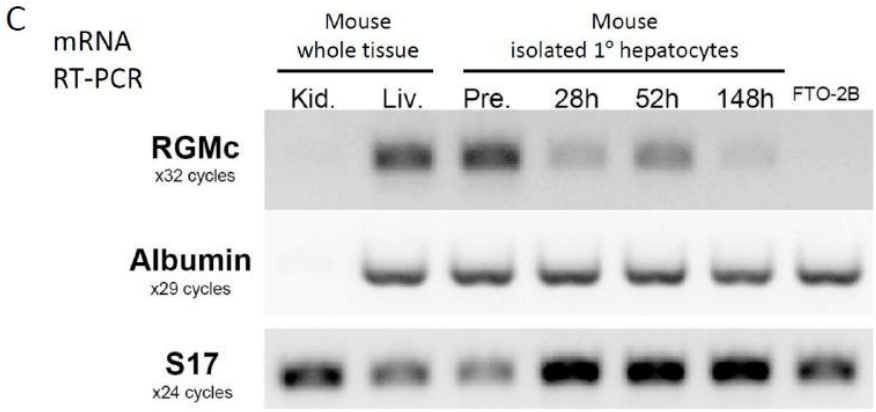
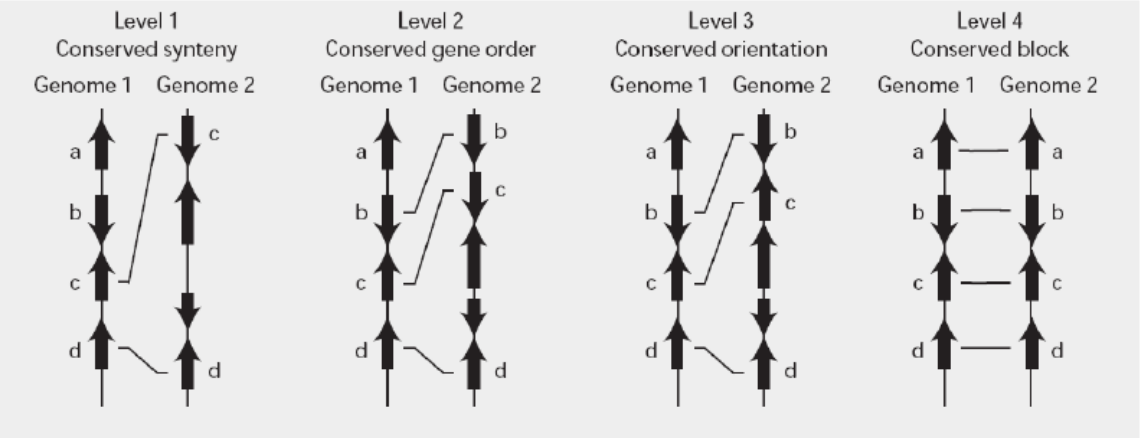

C
mRNA
RT-PCR
Mouse
whole tissue
Mouse
isolated 1° hepatocytes
Kid.
Liv.
Pre.
28h
52h
148h
FTO-2B
RGMc
x32 cycles
Albumin
x29 cycles
S17
x24 cycles

Figure 3.20. **Concept of gene and promoter synteny**. Levels of genomic conservation. Figure adapted and taken from Ref. [36] (Catchen, 2008). Levels 1 through 4 represent increasing amounts of conserved synteny. Arrows represent genes on a chromosome, or alternatively promoter elements in front of a gene which may or may not be orientation dependent. Level 1 requires only that two orthologous genes (or promoter elements) occur on homologous chromosomes/regions, level 2 requires conserved gene/element order; level 3 additionally requires conserved transcription orientation, and level 4 requires no intervening genes/elements (or element orientation if applicable to a transcription factor binding site) within the conserved block.

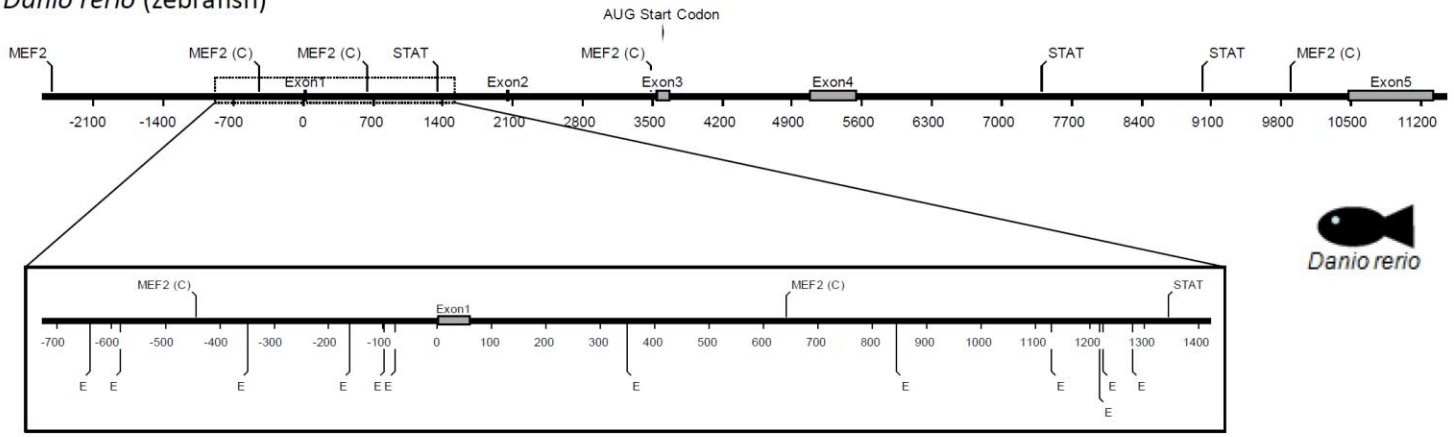
“Promoter Synteny”



Inferring Ancestral Chromosomes
J. Catchen, J. Conery, J. Postlethwait
Univ. of Oregon

Figure 3.21. **Possible promoter synteny within the RGMc locus with zebrafish and putative RGM in Ciona.** Putative transcription factor binding sites (TFBS), based on sequence alone (E-box: -CANNTG-; STAT: -TTCNNNGAA; MEF2: -BRMCWAWHRWRGBM- , where N is any nucleotide, B is not A, R is a purine, M is A or C, W is an A or T (weak), H is not G) in the RGMc locus of zebrafish (*Danio rerio*) and the RGM locus of sea squirt (*Ciona intesinalis*). MEF2 canonical sequence derived from Refs. [150, 222]. Exons shown as *gray boxes* for reference.(Ciona photo courtesy of Arjan Gittenberger, ascidians.com)

Danio rerio (zebrafish)



Ciona intestinalis (sea squirt)

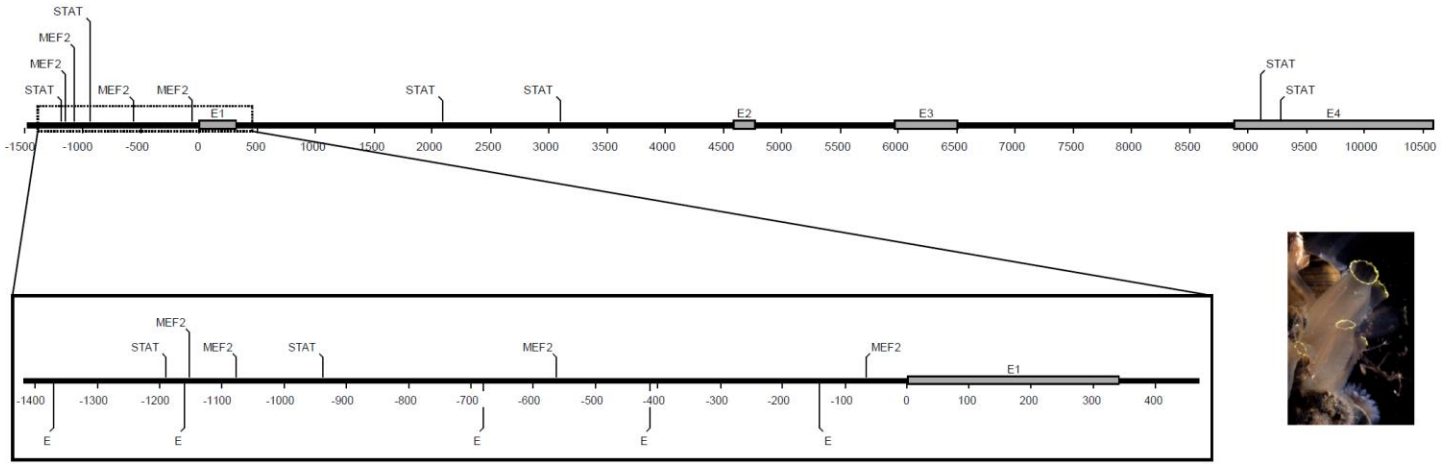


Figure 3.22. **Evolutionary analysis of dimerizing transcription factors.** Origin and repertoire of families of transcription factors that dimerize/multimerize in eukaryotes. The phylogentic tree represents the divergence of species based on molecular clock studies [223]. The transcription factor gene numbers, subfamilies, and their evolution are based on published studies (see [219], and references 7, 14-25 therein) with the following abbreviations used: bHLH, basic helix-loop-helix; bZIP, basic leucine zipper domain; NR, nuclear receptor family; MADS-box, MCM1(minichromosome maintenance)–agamous–deficiens–serum response factor; STAT, signal transducers and activators of transcription; NF-κB, nuclear factor kappa-light-chain-enhancer of activated B cells; HD-ZIP, homeodomain-leucine zipper (specific to plants). Figure adapted and taken from Ref. [219] (Amoutzias, 2008).**

** From Ref. (Amoutzias, TiBS (2008) v335:220-229) "The bZIP, bHLH and MADS-box families emerged at the origin of eukaryotes and are shared by plants, fungi and animals. The bZIPs and bHLHs underwent independent lineage-specific expansions in plants and animals. The MADS-box proteins underwent lineage-specific expansion in plants. Although no STATs have been identified so far in fungi, they are found in an opisthokont slime mould, Dictyostelium discoideum, and in animals. Nevertheless, this family did not undergo significant expansions. NRs emerged at the origin of animals and have undergone significant expansion, thus forming a complex dimerization network in humans. NF-kB is animal specific but did not undergo significant expansions. HD-ZIPs are plant specific and the family expanded significantly. In conclusion, five families expanded independently and significantly in animals and/or plants but not in fungi."

	Homo sapiens		Saccharomyces cerevisiae		Arabidopsis thaliana	
	Genes	Subfamilies	Genes	Subfamilies	Genes	Subfamilies
bHLH	118	44	8	?	162	12-21
bZIP	51	19	13-14	6-7	75	10
NR	48	19	-	-	-	-
MADS-Box	5	2	4	2	107	5
STAT	7	2	-	-	-	-
NF-κB	5	2	-	-	-	-
HD-ZIP	-	-	-	-	47	4

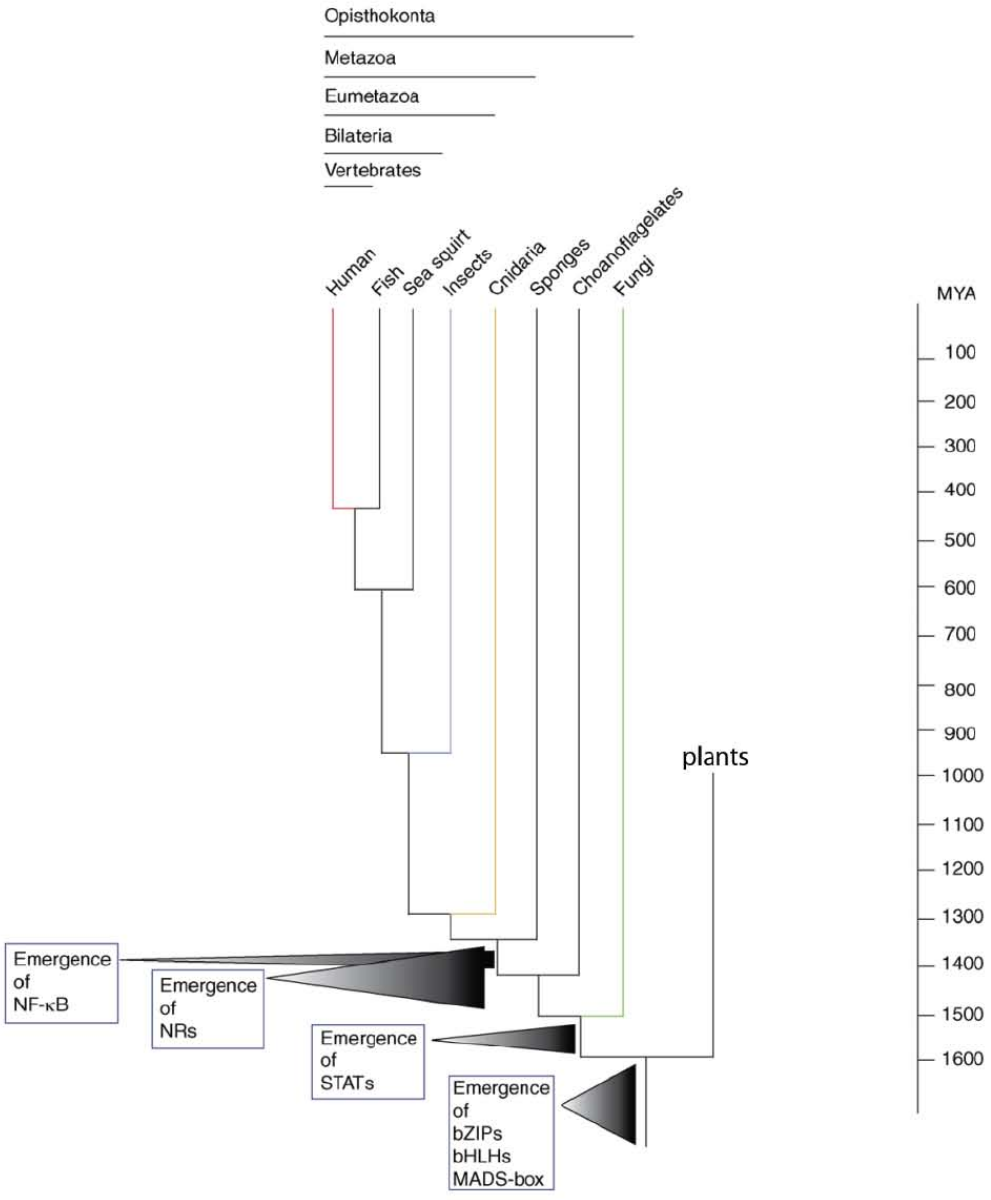
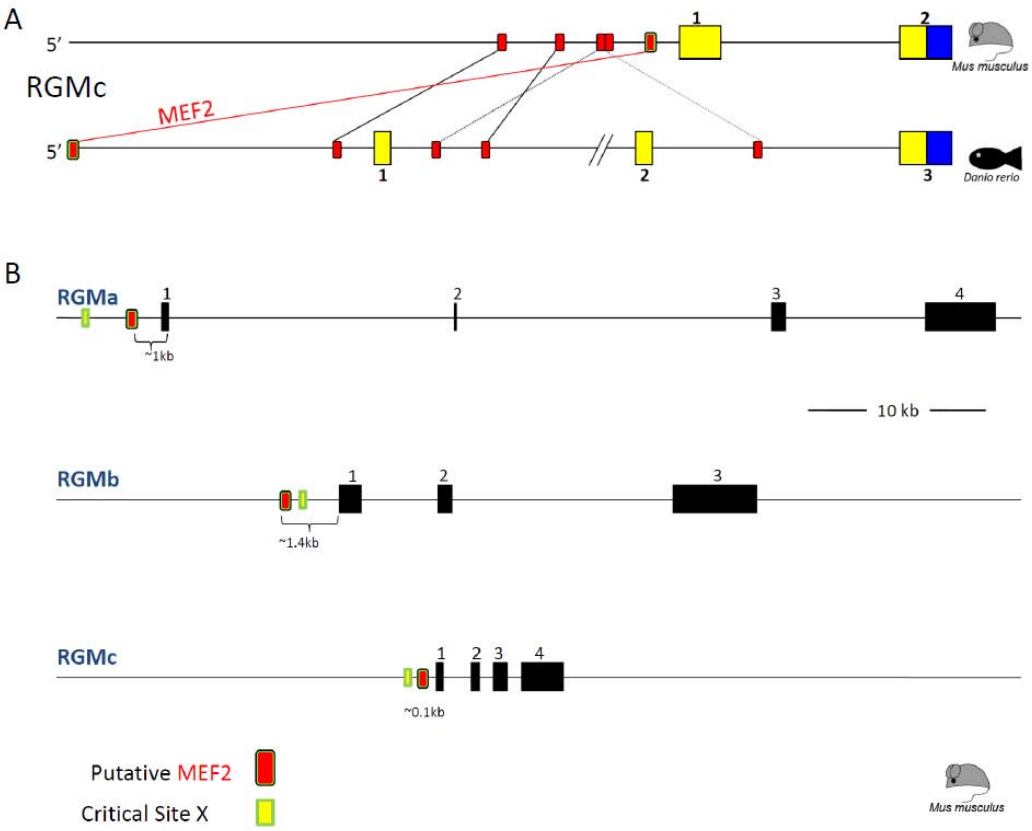


Figure 3.23. **Synteny across the promoters of the RGM family.** In classical genetics, synteny is the presence of two genetic loci on the same chromosome; in modern molecular genetics, synteny is defined as regions of large conserved blocks of DNA that would be expected to undergo rearrangement under random breakage model of chromosome evolution [224]. The same concept is extended here to promoter elements (see Fig. 3.20) of the RGM family. **A.** Conserved regions of DNA (shown here as the same size in *red* to simplify diagram for illustrative purposes; actual sizes and orientation are highly variable) between mouse and zebrafish RGMc, with MEF2 site noted (*red with green outline*). Exons are numbered and shown as *boxes,*with untranslated regions in *yellow,* and protein coding sequences in *blue*. **B.** All three RGM family members in mouse contain a putative MEF2 site (*red with green outline*) based on sequence alone (see figure legend 3.21 for sequence), although the MEF2 site found in RGMa and RGMb are ~1 kb from the predicted transcription start site; the functional relevance of these sites in RGMa and RGMb are unknown, as neither of their repective promoters have been characterized to date. Hypothetical transcription factors binding sites that confer tissue-restricted expression of the individual RGM family members are shown in *yellow with a green outline*. Theoretically, this model supports the concept of 'Modular Gene Organization' [15-17] in which the ancestral RGM contains a MEF2 site that is bound and regulated by an evolutionarily ancient MADS-box transcription factor (see Fig. 3.22 and Ref. [219]). Following gene duplication, the individual RGM family members may have used an additional binding site (*yellow/green*) in addition to the MEF2 site to confer the tissue-restricted pattern of expression seen in modern vertebrates that express the three RGM family members.


A
5′
RGMc
MEF2
5′
1
2
Mus musculus
1
2
3
Danio rerio
B
RGMa
1
2
3
4
~1kb
10 kb
RGMb
1
2
3
~1.4kb
RGMc
1
2
3
4
~0.1kb
Putative MEF2
Critical Site X
Mus musculus

(This page was intentionally left blank)

Chapter 4

Post-Transcriptional regulatory mechanisms of RGMc expression

"Excellence in research requires extended, undisturbed time and space in which to develop, evaluate, and fully test an idea. It demands the freedom to fail and try again." –Amy C. Justice, 2009

"I would rather discover a single fact, even a small one, than debate the great issues at length without discovering anything at all." –Galileo Galilei

The research presented in this chapter is being prepared for publication as:

Repulsive Guidance Molecule c / Hemojuvelin regulation by a post-transcriptional element in the 5'-untranslated region.

Christopher J. Severyn and Peter Rotwein

Department of Biochemistry and Molecular Biology, Oregon Health & Science University,
3181 SW Sam Jackson Park Road, Portland, OR 97239-3098, U.S.A.

In preparation June 2010

This work was supported by the National Institutes of Health,
grant numbers R01 DK42748 (to P.R.), T32 HL007781 (Training Grant in Molecular Hematology) and F30 HL095327 (to C. J.S.).

4.1: Summary

Transcriptional regulatory mechanisms of RGMc were presented in chapter 3 of this dissertation, however the full spectrum of RGMc gene regulation remains incomplete. In this fourth chapter, I define a novel post-transcriptional control region found within exon 1, called the ε-element, that increases reporter expression by 10-fold in muscle cells, and ~40-fold increase in three unique liver cell lines. As the levels of mRNA do not appreciably change with the addition of this region, the data presented in this chapter support the hypothesis that the ε-element is acting as a translational control element. The ε-element can function in both the forward and reverse orientation, but must be directly adjacent to the rest of exon 1. The 5'UTR of RGMc does not appear responsive to changes in iron levels, however alternative splice variants in exon 2 that change the size of the 5'UTR (presented in chapter 3) may slightly alter the levels of expression, potentially adding an additional control mechanism for the expression of the RGMc gene. Evolutionary conservation of the RGMc 5'UTR as well as experiments to discern possible control mechanisms are discussed at the end of the chapter.

4.2: Introduction

Comprehension of the possible structure-function relationships of repulsive guidance molecule c (RGMc), or hemojuvelin (HJV), at the protein level has improved greatly since its discovery, however there are still many gaps in our knowledge about the members of the RGM family (several of which are discussed in chapter 2) and the steps before the appearance of protein (the focus of chapters 3 and 4). Understanding the molecular mechanisms responsible for control of RGMc biosynthesis under physiological

or pathological conditions has been hampered by lack of fundamental information about the RGMc gene. In the preceding chapters, we defined the structure, examined the evolution, and established regulatory mechanisms of the mouse RGMc gene at the level of transcription. In skeletal muscle, three conserved elements were identified, that when collectively mutated, abrogated all apparent transcriptional activity in two muscle cell culture systems. Furthermore, RGMc appears to be contolled by basic helix-loop-helix and MEF2 transcription factor families in skeletal muscle, and potentially in hepatocytes by the liver-enriched transcription factor HNF4α (see chapter 3). In this chapter, a novel post-transcriptional control element is presented, and insights into the possible mechanisms of RGMc regulation following the appearance of a transcript, but before the presence of a nascent protein, are examined. Furthermore, the control element in RGMc may be an example of a small, but growing number of regulatory mechanisms that utilize the 5'-untranslated region (UTR) to enhance translation of specific mRNA transcripts into a nascent protein.

4.3: Experimental Procedures

4.3.1: Materials – Detailed in section 3.3.1 along with the following additions: anti-human ferritin (F5012) antibody, deferroxamine mesylate (DFO, D9533), ferric (iron III) ammonium citrate (FAC, F5879) were from Sigma (St. Louis, MO). Oligonucleotides were synthesized at the OHSU DNA Services Core. All other chemicals were reagent grade and were purchased from commercial suppliers.

4.3.2: Construction of RGMc promoter-reporter plasmids – Mouse RGMc genomic DNA was isolated from BAC clone RP24-136I19 (Children's Hospital Oakland Research

Institute BACPAC resource center (http://bacpac.chori.org/), Oakland, CA). Additional details may be found in section 3.3.2.

4.3.3: Cell culture and recombinant adenoviruses – Identical to methods listed in sections 3.3.3-.4. For iron-loading experiments, DFO and FAC were resuspended in the appropriate growth media (listed in section 3.3.3) at a final concentration of 500 μM.

4.3.4: Analysis of gene transcription by promoter-reporter gene assays - C2 and 10T½ on gelatin-coated 12-well plates and treated as noted in section 3.3.5. Liver cells (*Hep3B, HepG2,* and *AML-12*) on 12-well plates were transfected at 50% confluent density with 0.5 μg of plasmid DNA per well for 24 hours. Liver cells were then transferred into growth media (see section 3.3.3 for specifics for each cell type) for an additional 24 hours and then cell extracts were harvested, stored at -80^{o}C, and were assayed together for luciferase activity, as described [136], and results were normalized to cellular protein concentrations.

4.3.5: RNA isolation and analysis - Total cellular RNA was isolated from cells and tissues using Trizol as noted in section 3.3.7. Sequences of additional primers (depicted in Fig. 4.4D) targeted to the luciferase gene for steady-state mRNA analysis are as follows: Forward #1, 5' GCTTACTGGGACGAAGACGAACA; Forward #2, 5' CGATGACGGAAAAAGAGATCGTG; Common Reverse Primer, 5' GCATTCTAGTTGTGGTTTGTCCAAACTC.

4.3.6: Protein isolation and immunoblotting - Whole cell protein lysates were prepared as described [131] and aliquots stored at -80°C until use. Protein samples (25 μg/lane) were separated by SDS-PAGE, transferred to Immobilon-FL, blocked in AquaBlock, and incubated with primary and secondary antibodies at the following dilutions: anti-human

ferritin rabbit polyclonal antibody (1:2500), anti-α-tubulin (1:30,000), and AlexaFluor 680-conjugated goat anti-mouse IgG or IR800-conjugated goat anti-rabbit IgG (1:5000). Immunoblot images were acquired using a LiCoR Odyssey Infrared Imaging System, and analyzed with v2.0 analysis software (LiCoR, Lincoln, NE).

4.3.7: Data Analysis and Computational (in silico) Resources – Identical to methods listed in sections 3.3.12-.13, with the following addition: WebLogos@Berkeley [225, 226].

4.4: Results

4.4.1: RGMc contains a post-transcriptional control element within exon 1.

As demonstrated in chapter 3, analysis of a ~12 kb locus demonstrates that the major transcriptional regulatory elements for RGMc in skeletal muscle are found within a 0.6 kb region just upstream of the transcription start site (Figs. 3.8 through 3.14). In particular, data shown in Fig 3.12 revealed that well-conserved regions between mouse and human RGMc/HJV contained no apparent transcriptional enhancer (or repressor) elements (defined as regions that can alter transcriptional activity when placed at a distance and any orientation from a promoter element). During this analysis, it became apparent that while a discontinuous stretch of the RGMc locus did not alter luciferase-reporter activity (Fig. 3.12), a region of the RGMc locus continuous with the promoter region increased reporter activity by an order of magnitude (Fig. 4.1B, note *logarithmic* scale). Subsequent 3' deletions revealed that a region of only 42 bps, from +118 to +160, was sufficient to cause a 10-fold increase in luciferase activity of the RGMc promoter-reporter construct (Fig. 4.1B), identified in chapter 3 as a region located from -620 to

+118 relative to the transcription start site (Fig 3.10). Importantly, the 10-fold increase occurred at both t0 and t24 or t48 hours after differentiation for Ad:MyoD-10T½ and C2 cells, respectively (Fig. 4.1B), as well as when the three transcriptional control elements (α-, β-, and γ-elements, see Figs. 3.13 and 3.14) important for transcription in skeletal muscle were mutated individually and in combination (Fig 4.2 and appendix 1). This suggests that this 42-bp element operates independently of the known transcriptional response elements. As this 42-bp consistently gave an 'enhanced' level of reporter activity, the region was named and will subsequently be called the epsilon-, 'ε-', element.

Identification of the 42-bp ε-element suggested multiple potential mechanisms of enhanced luciferase-reporter activity. As diagrammed in figure 4.3, three possible mechanisms included (i) transcriptional, (ii) variable mRNA stability, and (iii) translational regulation to alter the expression of luciferase driven by the RGMc promoter; this list is by no means exhaustive, but intended to provide a framework of testable hypotheses from which experiments may be designed. For example, if the increase in reporter-activity is due to a transcriptional element being bound by a transcription factor present in proliferating and differentiating myoblasts (Fig. 4.3A), one would expect a concomitant increase in mRNA and reporter-activity (as a surrogate for protein levels). At the same time, an alternative hypothesis that would be supported by the same data set could be altered stability of the mRNA transcript (Fig. 4.3C, *right*), which effectively increases the total mRNA available for translation. A third possibility is changes in the translatability of the mRNA (Fig. 4.3B, *left*), often due to structural changes in the mRNA and/or proteins that bind to the ribonucleic acid. To distinguish between the mechanisms of translatability versus transcription or mRNA stability, we isolated total mRNA from differentiating myoblasts and compared total mRNA to the relative lucerase activity of RGMc promoter constructs with and without the 42-bp ε-

element (Fig. 4.1 C and D). As shown in figure 4.1D, there is no appreciable difference in the levels of total mRNA, as assessed by RT-PCR, whether the ε-element is absent or present (+118 *vs.* +160, Fig. 4.1D), while the reporter activity increases by approximately 10-fold (Fig. 1C, first two constructs). Together, these data are consistent with the hypothesis that the ε-element is changing luciferase-reporter levels by a translational control mechanism, and are not compatible with models of transcriptional control, nor changes in mRNA stability.

4.4.2: The RGMc ε-element functions regardless of orientation, but must be adjacent to the rest of exon 1.

To determine the nature of the ε-element, a series of reporter constructs were generated (diagrammed in Fig 4.1C, *left*) and subsequently tested in the Ad:MyoD-10T½ myoblast model. Reversing the orientation of the ε-element did not alter the 10-fold increase in reporter activity over the RGMc promoter alone (compare the first three constructs in Fig. 4.1C), but inserting the luciferase gene between +1/+118 of exon 1 and the ε-element (+118/+160) caused the luciferase activity to return to the levels seen with the RGMc promoter alone (Fig. 4.1C). Of note, transferring the ε-element to a heterologous promoter, both TK-Luc and myogenin-Luc, caused a decrease in activity (data not shown), although this could be attributed to the manner in which the reporter plasmids were constructed. Together these data in figures 4.1 and 4.2 demonstrate that the RGMc ε-element is a post-transcriptional regulatory element that can boost reporter activity by 10-fold when placed in any orientation, but must be adjacent to the rest of 5'UTR in the RGMc transcript.

4.4.3: Alternative splicing in RGMc exon 2 leads to three 5'UTRs of different size.

Initial mapping of the RGMc transcripts (presented in chapter 3) revealed a variable splice-acceptor site in exon 2 that allows the generation of thee different sizes of the

5'UTR (Figs. 3.2 and 3.3), but does not alter the protein coding sequence (CDS). To determine if this alternative splicing was unique to skeletal muscle, we tested a variety of tissues and cell lines known to express RGMc using RT-PCR with primers targeted to amplify across the exon 1-2 junction (Fig. 4.4A, *right*). When amplified against a sequence-verified positive control, all tissues and cell lines that express RGMc appear to contain the three splice variants (Fig. 4.1A). The 5'UTR in RGMc is larger than average when compared to the size of other eukaryotic UTRs (Fig. 3.4), which is often correlated with additional regulatory elements for gene expression at the level of translation.

4.4.4: RGMc ε-element increases reporter activity irrespective of the 5'UTR size.

To determine if the three different-sized 5'UTRs altered reporter-expression, three additional constructs with the endogenous 5'UTRs of RGMc were constructed, varying in size from 323 to 246 nucleotides (Fig 4.4B). Ad:MyoD-10T½ cells were transfected with five different fragments of the 5'UTR (called omega, 'Ω', plus the UTR size in nucleotides) and assayed for luciferase-activity as well as total mRNA with primers directed against the 3'end of the luciferase gene (diagrammed in Fig. 4.4D) in order to assess reporter-activity and mRNA levels in the same experiment. As shown in figure 4.4C, all constructs that include the 42-bp ε-element in exon 1 increase reporter activity by 5 to 10-fold over the RGMc promoter (Ω-118 in Fig. 4.4C) in differentiating muscle cells. Furthermore, levels of mRNA are relatively equal (Fig. 4.4E), further supporting the hypothesis that the order of magnitude change in reporter activity is not due to changes in mRNA levels, but utilization of the transcript (e.g., translatability of the mRNA). The difference in reporter activity with the Ω-323 construct (Fig. 4.4C) will be discussed in detail below, as the decrease could be attributed to mRNA levels (Fig. 4.4E) or unique features (e.g., sequence composition and possible secondary structures) that alter the levels of expression (Figs. 4.5, 4.9, and 4.10).

4.4.5: The RGMc 5'UTR and ε-element in liver cells, and under changes in iron levels.

RGMc may be one of the only genes expressed exclusively in striated muscle and in hepatocytes. As the liver is often recognized as an important regulator of iron metabolism, we sought to investigate RGMc promoter function in liver cells and thus performed a series of transfection experiments with mouse RGMc promoter-luciferase fusion genes into human Hep3B liver hepatoma cells, which have been reported to produce RGMc (mRNA) [72], as well as into mouse AML-12 hepatocytes and human HepG2 hepatocarcinoma cells. None of the 5 promoter deletions tested were active, as luciferase values were < 10% of what was measured with a thymidine kinase promoter-reporter plasmid, and were < 1% of the activity seen with mouse Igf-II promoter 3 (Fig. 3.18). Similarly negative results were seen when the region from +2812 to +7540 of the mouse RGMc gene was added to the promoter-reporter plasmid (data not shown), despite the region being highly conserved between mouse and human (Fig. 3.12A). Based on these results, it appears that the regulatory domains responsible for RGMc transcriptional activity in liver cells do not map to the promoter of the gene, however as noted in chapter 3, the transcription factor HNF4α appears to be important for RGMc expression as shown by ChIP-Seq in human cells and mouse knock-out studies of HNF4α that show no expression of RGMc. Alternatively, the transcription factors necessary for RGMc expression may not produced in Hep3B, AML-12, or HepG2 cells. Interestingly, when the 42-bp ε-element in exon 1 is included, the reporter activity increases ~40-fold (as a note, levels increase from a virtually undetectable signal to a level ~2 times greater than thymidine kinase) in Hep3B cells (Fig. 4.7, *gray hatched*, -620/+1320 construct), with similar results in HepG2 and AML-12 cells (Fig. 4.7) suggesting that either (i) the transcriptional elements necessary for RGMc expression can be found within this region and/or (ii) an additional post-transcriptional regulatory mechanism (e.g., translational control) is operating on this region.

To determine if 5'UTR alters RGMc expression, the five different (three endogenously derived, two synthetic) 5'UTR-luciferase constructs were transfected into Hep3B cells, and mRNA levels were assayed in parallel as described above. As shown in figure 4.8, simply adding the ε-element increases reporter activity (Fig. 4.8A, lanes Ω-160 *vs.* Ω-118), without any concomitant increase in mRNA levels (Fig. 4.8B), a similar trend to what was seen in the skeletal muscle system. In contrast, the Ω-323, Ω-264, and Ω-246 constructs all showed a 50% reduction in reporter activity as the when compared to Ω-160 (Fig. 4.8), whereas only the Ω-323 showed a reduction in the skeletal muscle Fig. 4.4C). Possible reasons for this are discussed below, however most importantly, the addition of the ε-element confers a dramatic increase in reporter activity regardless of the size of the 5'UTR in Hep3B cells.

Since RGMc is an important component of an iron regulatory pathway, and UTRs have been shown to have iron responsive elements (Refs. [129, 227] and see Fig. 4.3), we repeated the transfection experiments with the varying sizes of the 5'UTR in Hep3B cells under iron-depleted and iron overloaded conditions using the iron chelator deferroxamine (DFO) and an iron delivery system ferric (iron III) ammonium citrate (FAC), respectively (Fig. 4.8C). The 5'UTR of RGMc does not appear to be responsive to changes in iron levels (Fig. 4.8C). As a control, we were able to demonstrate that the cells were appropriately iron depleted and iron loaded via a western blot of ferritin (Fig. 4.8D), a well-characterized protein shown to change its protein levels in response to iron levels [129].

4.5: Discussion

Experiments in this chapter demonstrate that a novel regulatory element in the 5'UTR of RGMc controls gene expression by a post-transcriptional mechanism. This regulatory element, called the ε-element, localizes to a 42-bp region exon 1, and while it must be directly adjacent to the rest of exon 1, the orientation of the ε-element does not appear to matter (Fig. 4.1C). As there is no appreciable change in the mRNA levels (Figs. 4.1D, 4.4E, and 4.8B) with an order of magnitude increase in reporter activity, a translational regulatory mechanism appears to be the most parsimonious explanation for these data. In this discussion, I will outline the possible mechanisms by which the the ε-element could be regulating RGMc expression, integrating the data presented in this chapter, previously characterized examples of post-transcriptional regulation that may pertain to RGMc, and computational data on the 5'UTR of RGMc that may shed light onto the nature of the regulatory mechanisms of RGMc expression.

The central concept in translational control is that gene expression is regulated by the efficiency of *how the mRNA is utilized* in specifying the synthesis of a nascent protein [22]. As shown in figure 4.3, three hypotheses regarding the possible nature of the ε-element were proposed. The models for transcriptional regulation and mRNA stability do not appear to be consistent with the available evidence. Data in this chapter support the hypothesis that the ε-element is regulated via translational control, which has many examples, but are limited in numbers of detailed molecular mechanisms [228]. One may envision that our understanding of translational regulation by proteins and ribonuclear proteins (RNP) will rapidly become similar to the detailed understanding of transcriptional control that is emerging (with the intricate interplay between the DNA, chromatin, transcription and chromatin modifying machinery), and that the mRNA/mRNP of translation will unfold as a model in which regulation occus via a

"chromatinziation of mRNA" [229]. Part of this understanding has been accelerated by the use of genomic data and molecular evolution to develop hypotheses to experiments that nature has already performed on entire organisms. Below are possible mechansisms that may influence how the mRNA is used, how this utilization ultimately may influence the expression of RGMc, and how these principles may be applied to our understanding of the molecular evolution of a family of genes.

Alternative Splicing: changing the scaffold and interaction matrix

Before an mRNA is translated into a protein, the newly synthesized transcript must be processed for export to the cytoplasm. This includes a series of steps that usally includes capping the 5' end, adding a poly-adenylation tail, and generating alternatively spliced transcripts as appropriate. Estimates of the number of genes that undergo alternative splicing range anywhere between a third (34% within genes of the endocrine systems, reported in 2001) [230], to 50% [231] (though there are errors in this computationally derived database released in 2009, including RGMc being mis-annotated), to a recent high-throughput sequencing study suggesting that as many as 95–100% of human pre-mRNAs *with more than one exon* are processed to yield multiple mRNAs [187, 232, 233]. Alternative splicing is probably of particular importance for developmental and stage-specific isoforms [230] and it has been shown that threre are multiple changes in the 5'UTR that may influence expression (reviewed in Refs. [187, 234]).

As RGMc has three splice variants that are all longer than average (Fig. 3.4) and well-conserved (Fig. 3.8), it is likely that this 5'UTR contains important regulatory elements for RGMc expression. Shown in chapter 3 (see Figs. 3.2C and 3.3), the three alternatively spliced exon-2 acceptor sites are well-conserved in mammals, with a canonical poly-pyrimidine tract (also see Fig. 4.5A, *green,* % C+T) and –AG splice acceptor site (additional details of sequence composition will be discussed below). While

all three splice variants appear present in all cell lines and tissues that express RGMc (Fig 4.4A), the function of these variants remains unknown, as the change in reporter levels only drop by ~2-fold (Figs. 4.4C and 4.8A), and are not affected by changes in iron levels (Fig. 4.8C). It will be interesting to see if the ratio of splice variants changes during the developing embryo or under pathological conditions, and if the regulation of the alternative splice-variants is at the translational level (via output of a reporter or protein levels).

When investigating possible translational mechanisms, it is important to rule out that simply removing the intron, which can act as an inhibitory sequence from the 5'UTR, is not the cause of increased reporter activity [235]. In the case of the RGMc reporter constructs, whether the intron is present (Fig. 4.1) or absent with a continuous stretch of 5'UTR sequence (Figs. 4.4 and 4.8), the levels of reporter activity have a 10-fold increase in activity when compared to the RGMc promoter-construct alone. This suggests that we may rule out that splicing out the intron is the reason for the increased reporter activity.

The 5'UTR: a functional hot-spot for translational regulation

Classically the untranslated regions (UTR) of mRNA are used for translational regulatory mechanisms. Iron regulation is governed, in part, by two canonical strategies with examples in (i) ferritin being regulated in the 5'UTR by blocking translation (inhibiting the recruitment of the 43S ribosomal complex to ferritin mRNA [236, 237]), and (ii) by the 3'UTR in transferrin acting to stabilize the mRNA and inhibit degradation (and thus effectively increasing the number of transcripts available for translation) (Fig. 4.3B and C). Both ferritin and transferrin are responsive to changes in iron levels [129] using the iron-regulatory protein *cis-aconitase* that binds an iron regulatory element (IRE), a secondary structure in the mRNA to which Aconitase/IRP (Iron Regulatory element binding Protein) binds and alters translation. The 5'UTR of RGMc is not responsive to

changing the levels of iron (Fig. 4.8C), nor do total mRNA levels change in mice loaded with iron [238]. Thus, collectively the current data only support a mechanism by which RGMc is responsive to iron loading at the level of protein release from the cell membrane [67], and not at the level of the transcript. Furthermore, no classical IREs can be found in RGMc, thus it is likely that increase in reporter activity with the ε-element occurs by an additional regulatory mechanism other than iron responsiveness. Several possible mechanisms include utilization of upstream open reading frames (uORF), proteins that bind specific structures or sequence motifs, and secondary structures of the RNA sequence that may result in hairpin or internal ribosome entry site (IRES) motifs, as well as emerging regulation by miRNAs that may occur at either the 5' or 3' UTR (Fig. 4.9), of which a select number will be discussed below.

Two elements that influence the translation of many mRNAs include the uORF and IRES sequences (Fig. 4.9). Translation in eukaryotes classically begins when the 43S 'pre-initiation' complex (consisting of a 40S ribosomal subunit, eukaryotic initiation factor 2 (eIF2)-GTP-Met-tRNA$^{Met}_{i}$ ternary complex (eIF2 TC), eIF3, eIF1, eIF1A, and probably eIF5) recognize the initiation codon, 'AUG' [228, 239-242]. Some mRNAs contain AUG-start codons upstream of the "principal-AUG" (pAUG) called uORFs, that may create a control mechanism for translation, change the N-terminus of the protein, or completely alter the reading frame [242]. Mouse RGMc contains 5 uORFs in addition to the pAUG (Fig. 4.5B, *arrowheads*). Three of the RGMc uORFs, plus the pAUG, are well-conserved among mammals and contain a Kozak Consensus Sequence (GCC(***A/G***)CCAUG***G***) [-3 and +4 residues, highlighted in ***bold*** and ***italics*** are the most important]) [155, 243, 244], suggesting that if uORFs are used as a translational control mechanism in RGMc, these four are the most likely candidates (Fig. 4.5B, *asterisks* under the *arrowheads* and sequence alignment in Fig. 4.6). While there is at least one example of a non-AUG start codon in eukaryotes (a CUG-start codon [245]), I chose to

ignore these as there is no reliable way to predict non-AUG start codons [242]. Figure 4.6 shows the conservation sequence of the entire 5'UTR, as it has been shown that the sequence 5' of the first uORF (in yeast GCN4) is required for efficient translation [246], and thus may also be critical for RGMc if uORF are used as a control mechanism.

In addition, the spacing between the ATG/AUGs in RGMc are shown in figure 4.5B, as spacing between AUGs is important for "ribosomal shunting," a process by which the ribosome re-engages with the mRNA downstream of a translational block, with the classical example being found in Couliflower Mosaic Virus 35S mRNA [247], as well as for mRNA with IRES, which occurs via 5'-cap-*independent* mechanism [244, 247]. While RGMc does not have any predicted IRES, it is mentioned for completeness and the fact that RGMc has an unusually high percentage of pyrimidine enrichment (Fig. 4.5A) which is important for IRES sequences (predicting an IRES can be challenging as an IRES usually consists of a secondary structure that is not easily predicted, followed by a polypyrimidine tract [248, 249]), as well as for polypyrimidine tract binding proteins (PTB) which will be discussed below.

The next step in understanding the possible regulation of RGMc by uORFs would be to directly mutate the ATG/AUGs, beginning with those that contain a strong Kozak Consensus Sequence and are well conserved (Figs. 4.5B and 4.6), and determinine experimentally if they have a direct impact on RGMc expression. With respect to IRES translational control, searching for a protein that interacts with the 5'UTR of RGMc such as the IRES *trans*-Acting Factors (ItAFs) [248] is a possible way of determining the presence of an IRES if coupled with the 'gold-standard' bicistroinc assay used in cell culture [248]. Alternatively, a 'yeast three hybrid' [250] or other RNA-protein interaction screen may illuminate the various factors that control the 5'UTR of RGMc,

and its subsequent expression. A great deal of experimental work is needed to determine the precise mechanisms of regulation of RGMc in the 5'UTR sequence.

The RGMc 5'UTR contains a high percentage of pyrimidines within some of the exon 2 vaiants (Fig. 4.5A), as well as a large number of pyrimidines in the terminal 5' end of the transcript (the first 16/22 nucleotides (73%) are pyrimidines; see Fig. 4.6). While the ε-element appears to be the major post-transcriptional element in the RGMc 5'UTR (Figs. 4.4 and 4.8), the longer 5'UTRs of RGMc potentially play a second role in regulating RGMc expression. In muscle cells, the longest 5'UTR of 323 nucleotides (Fig. 4.4C, Ω-323) and three of the 5'UTRs in liver (Fig. 4.8A, Ω-323, -264, and -246) all show a modest, but reproducible (~2-fold) drop in activity when compare to Ω-160 construct. One possible explanation is the formation of a secondary structure (e.g., hairpin, Fig. 4.9, Table 4.1, and discussed below), while another is that the polypyrimidine-enriched region of RGMc (Fig. 4.5A) is bound by a protein such as the polypyrimidine tract-binding protein (PTB) [239, 251, 252]. Alternatively, the extreme 5' end of RGMc is enriched for pyrimidines, a feature that is often seen with many ribosomal proteins that contain a terminal 5'-oligopyrimidine (TOP) tract [241]. The TOP motif must be at the 5' end of the mRNA and may not always be sufficient without the remainder of the 5'UTR [241], however the mechanism remains unknown at this time. It is possible, that the ε-element itself requires the 5'end, though the orientation independence (Figure 4.1C) adds an additional complexity to understanding the mechanism of the ε-element along with the additional decrease in activity seen with the different 5'UTR constructs.

Possibility of a positive translational mRNA-specific regulator

Translational regulation appears to occur at either a global level, i.e., where large numbers of genes are affected, or in a gene-specific manner, such as those genes controlled by the iron regulatory protein [129]. The vast majority of regulatory elements

in the 5'UTR are of an inhibitory nature [234] (see Fig. 4.9), which include miRNAs that are able to bind to the 5' UTR in addition to their classical mechanism of 3'UTR target destruction. Examples of *positive* mRNA-specific regulators of 5'-cap-dependent translation in eukaryotes have been reported [230, 253, 254], although definitive examples of physiologically relevant control are limited and have not been sufficiently demonstrated [22]. For example, a *cis*-element in a splice-variant of the 5'UTR of pre-proinsulin mRNA appears to increase translation in response to glucose [219, 243], but this splice variant is less than 1% of the abundant native mRNA in pancreatic cells [230]. While most miRNAs operate by inhibiting translation via 3'UTR recognition (Fig. 4.9), another report suggests that a miRNA binds to the 5'UTR of ribosomal mRNAs to enhance their translation, suggesting a general method by which mRNAs with a 5'TOP tract are regulated [253], but the mechanistic details remain obscure at this time. Thus with respect to RGMc, the ε-element may represent a novel *positive translational control element* found in the 5'UTR. As all RGMc reporter constructs that contain the ε-element increase in activity without concomitant changes in mRNA levels (Fig 4.1 and 4.8), the current data is consistent with the hypothesis that the ε-element is a positive translational control region found in the 5'UTR.

<u>The ribosomal landing pad on the 5'UTR</u>

Translation efficiency of eukaryotic mRNAs is primarily dictated by initiation [239-241, 255]. The analysis above centers around one fundamental concept: if the hypothesis that '*the RGMc 5'UTR influences the translational expression of the gene*' is correct, then the mRNA elements, such as the ε-element, are all likely to recruit the critical factors for translational initiation. In this way, the 5'UTR acts as a scaffold through which the ribo-protein subunits localize and initiate translation. RNA probing studies indicate that the mRNA-binding site of the 40S subunit may cover from 12 to 18 additional nucleotides

upstream of the 40S subunit and about 15 nucleotides downstream of the initiation codon, so that initiation complexes assembled on mRNAs with short 5' leaders may lack stabilizing interactions upstream of the initiation codon [255, 256]. This may be influenced greatly by secondary setructure in the 5'UTR, which may include hairpins in the mRNA to complex structures recognized by specific mRNA binding proteins (e.g., the IRP or the IRES motif; see Figs. 4.3 and 4.9). mRNAs with a high percentage of CG-nucleotides often stabilze structural elements due to the three hydrogen bonds that can form between the G and C. RGMc contains two regions of the 5'UTR that are modestly enriched for GC-residues; one around the second uORF and one around the principal AUG (Fig. 4.5A). However, the predicted secondary structure around these regions varies little between an ensemble of theoretical predictions in mFold [257, 258] when compared to changes in experimental results (Figs. 4.1, 4.4, and 4.8). For example, the regions around +80 (Fig 4.5) is a GC-rich region and predicted to contain a hairpin structure in most of the models of the RGMc 5'UTR Ω-118, -160, and -323, but is absent in the majority of the models of Ω-246 and -264 (select examples shown in Fig. 4.10). The only trend from the secondary structures (Fig 4.10) that correlates with the experimental data (Figs. 4.1 and 4.8) is a modest change in the predicted $\Delta G^{o}_{folding}$ per nucleotide in the longer 5'UTRs (Ω-246, -264, -323) shown in Table 4.1. More fundamental experiments will be needed to determine if the secondary structure and GC-content of the 5'UTR alters the expression of RGMc, and whether alternative splicing creates a differential kinetic or thermodynamic requirement on the translation system [187] (the above analysis with ΔG^{o} is primarily a thermodynamic argument). For example, mRNAs containing extensive secondary structure in their 5'UTR translate efficiently in cells overexpressing initiation factor eIF-4E [259]. Additional computationally predicted motifs are listed in Table 4.2, with the most intruiging being an interferon-Gamma-Activated-Inhibitor of Translation (GAIT), and a weak 15-lipoxygenase Differentiation Control Element (LOX-DICE) [22, 260-262]. While

interesting from a speculative level about the fundamentals of RGMc regulation, greater experimental evidence will be needed before a mechanistic understanding of the role of the secondary structure of the RGMc 5'UTR may be understood.

4.6: Acknowledgements

We thank Bill Skach for his discussions and advice on the experiments presented in this chaper, as well as Lisa Wilson and David Kuninger for assistance with preliminary experiments, and other members of our laboratory for helpful comments during the development of the work presented in this chapter. The studies reported here have been supported in part by National Institutes of Health grants T32 HL007781 (Molecular Hematology Training Grant) and F30 HL095327 (to C. J. S.), and by R01 DK042748-21 (to P. R.).

Table 4.1: ***Predicted*** **ΔG^o** ***values for the folding of the RGMc 5'UTR***

RGMc 5'UTR Construct	Ω-118	Ω-160	Ω-246	Ω-264	Ω-323
Avergage ΔG^o	-30.1	-40.8	-68.5	-71.4	-90.4
ΔG^o Avg/nt	-0.255	-0.255	-0.278	-0.270	-0.279
(Number of structures analyzed in mFold)	(7)	(12)	(13)	(21)	(10)

Figure 4.1: RGMc contains a post-transcriptional control element within exon 1. A. Organization of the mouse RGMc gene. The gene contains 4 exons (*numbered boxes*) and three introns (*thin lines*), with the 5' untranslated region (UTR) in *gray* and the protein coding region in *black*. The transcription start site is denoted as a *bent arrow* with '+1'. **B.** Results of luciferase assays in differentiating Ad:MyoD-10T½ cells and C2 myoblasts that were transiently transfected with reporter genes containing the mouse myogenin promoter (as a control for muscle differentiation; data shown in figure 3.9), or 3' deletions of the mouse RGMc genomic DNA, which includes the RGMc promoter defined in chapter 3, (coordinates -620 to location listed adjacent to each construct), and incubated in differentiation media (DM) for 0 (*white bars*), or 24 or 48 hr (Ad:MyoD-10T and C2, respectively, *gray bars*). The graphs summarize results of ≥3 independent experiments (mean ± S.E.), each performed in duplicate. Myogenin promoter values at t 0 have been set to 100 in each graph (average measurements at t 0 were 7.8×10^4 (Ad-MyoD-10T½ cells) or 7.3×10^3 (C2 cells) relative light units/μg total protein/sec). **C**. Results of luciferase assays in Ad:MyoD-10T½ cells transfected with RGMc reporter constructs with different positions and oritentation of the RGMc ε-element (coordinates +118/+160) as diagramed to the left. For example, the third construct from the top places the ε-element in reverse orientation. Results are presented as in part **B**. Thymidine kinase (TK) and myogenin promoters are listed for reference. **D.** Total mRNA levels from Ad:MyoD-10T½ cells transfected with RGMc reporter constructs: -620/+118; -620/+160; REV, -620/+160 with reverse orientation of the ε-element at +118, and TK luciferase construct as a negative control. RT, reverse transcriptase added (+) or absent (-) from the reaction. Ribosomal S17 gene shown as a loading control. Scale Bar is the range of four independent experiments.

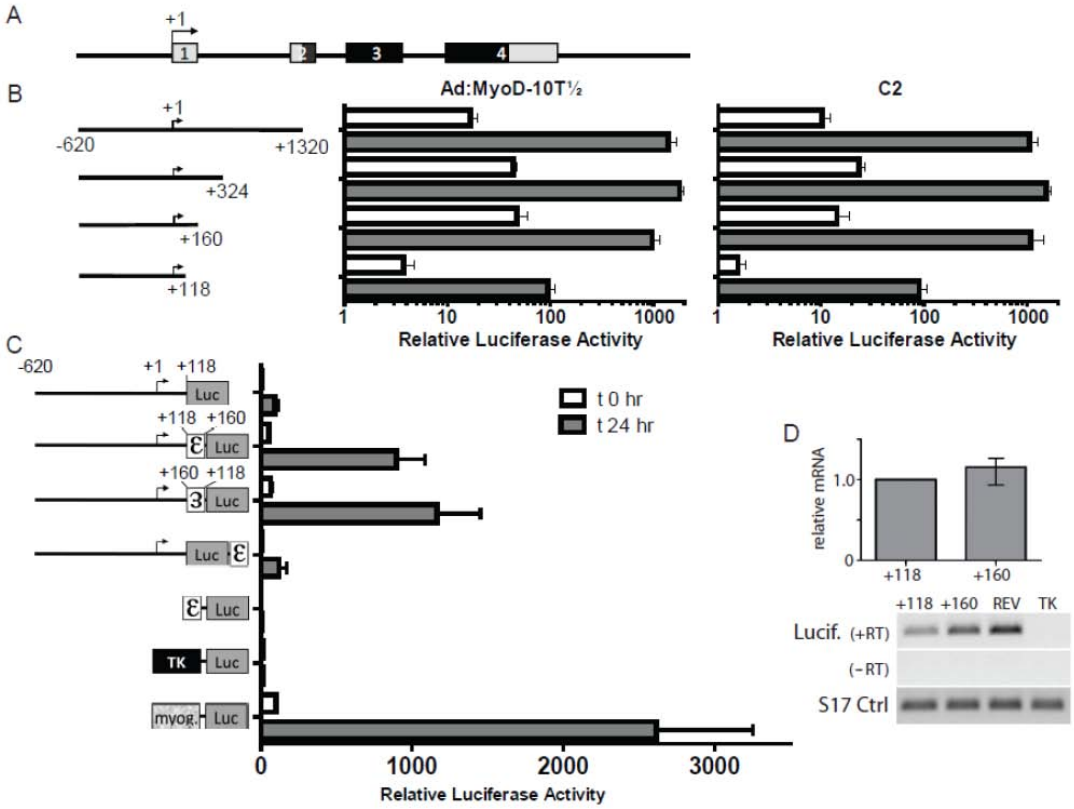

A
+1
1
2
3
4
B
Ad:MyoD-10T½
C2
+1
-620
+1320
+324
+160
+118
1
10
100
1000
Relative Luciferase Activity
C
-620
+1 +118
Luc
+118 +160
+160 +118
t 0 hr
t 24 hr
TK
myog.
0
1000
2000
3000
Relative Luciferase Activity
D
relative mRNA
1.0
0
+118
+160
+118 +160 REV TK
Lucif. (+RT)
(-RT)
S17 Ctrl

Figure 4.2: The RGMc ε-element does not appear to alter the promoter in skeletal muscle. Results are depicted of luciferase assays in differentiating Ad-MyoD-10T½ cells (*left panel*) and C2 myoblasts (*right panel*) transiently transfected with reporter genes containing substitution mutations of the mouse RGMc promoter (illustrated on the maps to the *far left*; details including full primer sequences are in '*Experimental Procedures'* of chapter 3). Figure demonstrates that the ε-element increases reporter activity by 10-fold irrespective of the time point (compare *top* figures without ε-element and *bottom* figures with the ε-element; note scale, with 10x increase on the lower graphs of constructs with the ε-element). Cells were incubated in DM for 0 (*white bars*), or 24 or 48 hr (*gray bars*) before analysis. The graphs depict results of 3 - 10 independent experiments (mean ± S.E.), each performed in duplicate (* - $p < 0.05$, ** - $p < 0.001$, vs. t 0). For full results of substitution mutations with and without the ε-element, please see appendix 1.

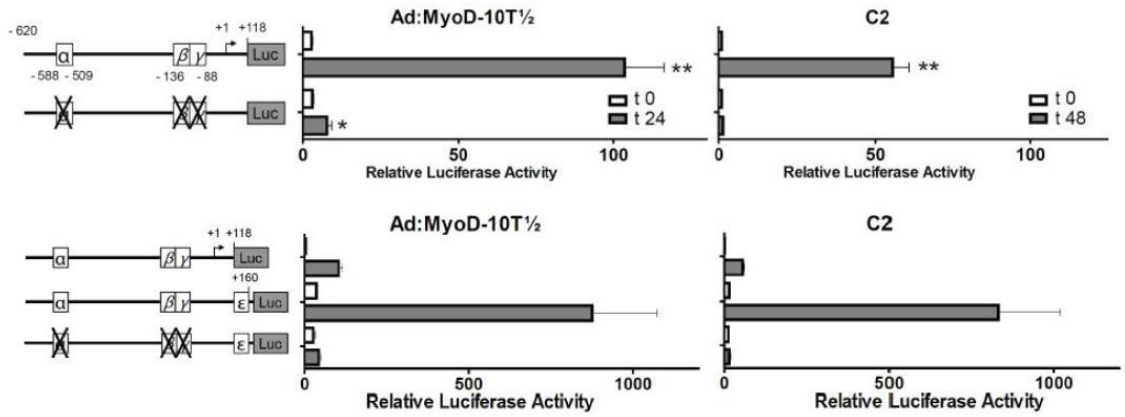

- 620
+1 +118
α
β γ
Luc
- 588 - 509
- 136 - 88
Ad:MyoD-10T½
C2
t 0
t 24
t 48
**
*
0
50
100
Relative Luciferase Activity
+160
ε
500
1000

Figure 4.3: Models of possible mechanisms of RGMc regulation by the ε-element. Proposed models to explain the increase in activity of the ε-element within the RGMc gene include a transcriptional control element, translational control and/or a mechanism of mRNA stability. The model of translational control appears most consistent with the data presented in this chapter, but as a novel 'activator element' instead of a translational repressor as depicted in **B** (please see text for details). **A.** Possible model of transcriptional regulation. A resident transcription factor (TF) labeled IV (*blue*) binds to the ε-element and activates transcription regardless of the status of the rest of the promoter, thus basal levels of transcription become enhanced by the presence of this factor. During muscle differentiation, the additional TFs interact cauing the RGMc promoter to become fully active. Examples of classical iron responsive elements, ferritin and transferrin receptor as examples of 5'UTR block in translation and 3' UTR regulation that stabilizes the mRNA from degradation. **B.** Ferritin is regulated by the iron regulatory protein (IRP) cytosolic-Aconitase which binds to the iron responsive element (IRE) in the 5'UTR when iron levels are low. Upon an increase in iron levels (Fe, *purple circle*), the IRP-Fe complex undergoes a structural change that no longer permits the binding of the IRP to the IRE, thus allowing the ferritin gene to be translated into a mature protein. **C.** The transferrin receptor 3'UTR, in contrast, operates to increase translation when no/low levels of iron is around by protecting the 3'UTR of the transferrin receptor transcript from nuclease degradation, effectively increasing the steaty-state levels of the transcript. **B** and **C** from Ref.[12], (Alberts, 2002).

Possible Mechanisms

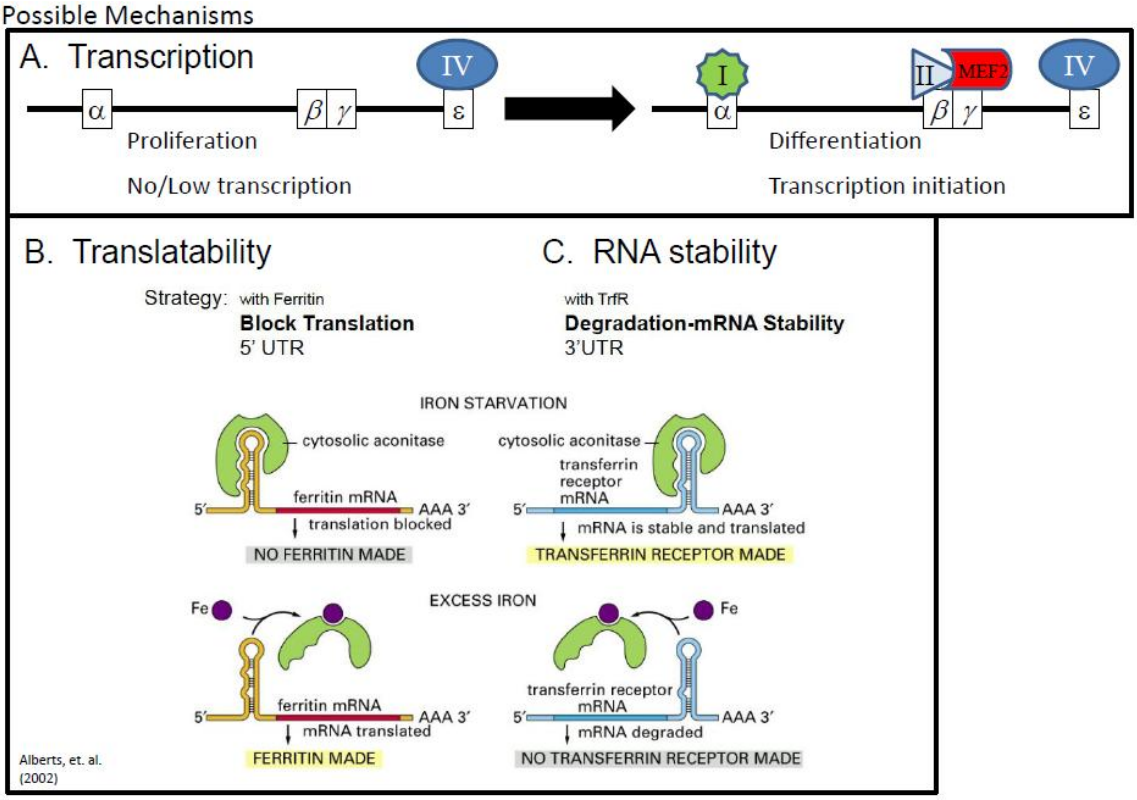


Figure 4.4: Alternative splicing in the 5'UTR of RGMc may alter translatability of the gene. Three alternatively spliced transcripts were characterized and discussed in Figs. 3.2 and 3.3. **A.** Results of RT-PCR amplification of mRNA using primers at the exon 1-exon 2 junction reveals that the three exon 2 splice variants are present in all cells tested that express RGMc. Samples include 10T½ cells transduced with Adenovirus expressing β-Gal (control not expressing RGMc, as shown in Ref. [34]) or MyoD at 24 hrs in DM; C2 cells at 48 hrs in DM; Mouse skeletal muscle (gastrocnemius), liver, heart, and control plasmids of sequence verified clones from skeletal muscle. Diagrams of the transcripts are shown to the *right* along with the length (in nucleotides) of the three exon 2 splice variants and length of the full 5'UTR. **B.** Restriction analysis of the RGMc reporter constructs in diagrammatic form (*left panel*), and cut with the restriction enzymes KasI, to demonstrate different sizes of the 5'UTR of the constructs (*center panel*) and NcoI (*right panel*) to linerize the plasmids and show equal concentrations for subsequent transfection into cells. **C.** Experimental set-up to determine the steady-state mRNA levels of the RGMc luciferase reporter (*yellow arrow*). Two unique forward primers were used (data in **D** the results from forward primer #1 [results from for. primer #2 were identical but shifted in size to the appropriate levels]). A common reverse primer localized to the nascent transcript just upstream of an SV-40 poly-Adenylation signal (*light blue*). Primer sequences may be found in *'experimental procedures'* of this chapter. **D.** Results of Ad:MyoD-10T½ cells transfected with RGMc reporter constructs and incubated in DM for 0 (*white bars*), or 24 hr (*gray bars*). Contructs are labeled as omega ('Ω') plus the size of the 5'UTR in nucleotides. The graphs summarize results of ≥3 independent experiments (mean ± S.E.), each performed in duplicate (* - $p < 0.02$, ** - $p < 0.5 \times 10^{-5}$). **E.** Data from an RT-PCR experiment of mRNA isolated at t24 hrs in DM to demonstrate approximatelt equal levels of mRNA expression from the different sized RGMc 5'UTRs.

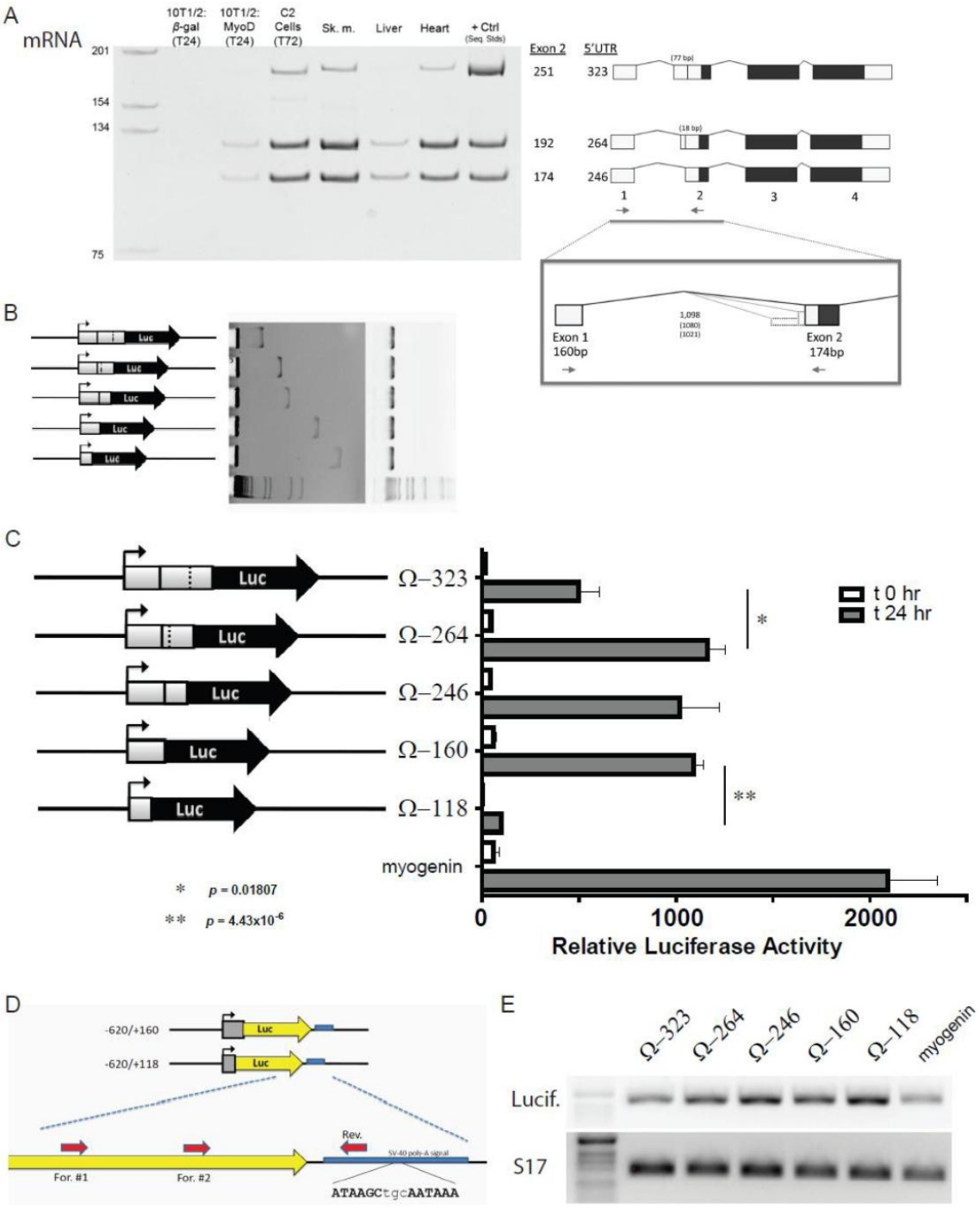
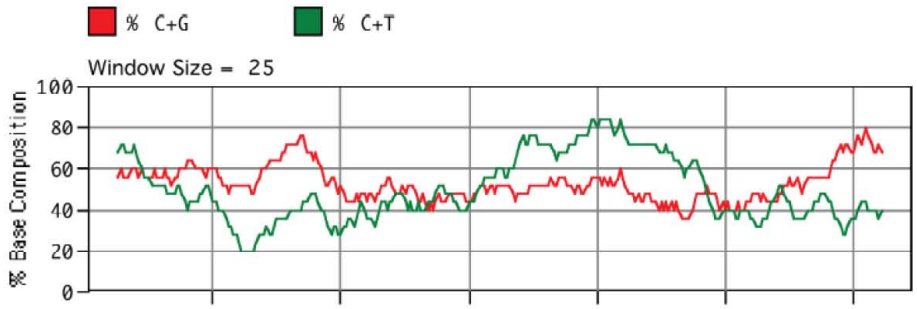
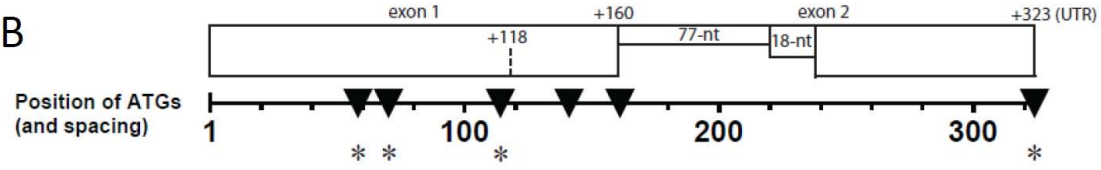

A
mRNA
10T1/2: β-gal (T24)
10T1/2: MyoD (T24)
C2 Cells (T72)
Sk. m.
Liver
Heart
+ Ctrl (Seq. Stds)
201
154
134
75
Exon 2
5'UTR
251
323
(77 bp)
192
264
(18 bp)
174
246
1
2
3
4
1,098
(1080)
(1021)
Exon 1
160bp
Exon 2
174bp
B
Luc
C
Ω–323
Ω–264
Ω–246
Ω–160
Ω–118
myogenin
t 0 hr
t 24 hr
*
**
* p = 0.01807
** p = 4.43x10-6
0
1000
2000
Relative Luciferase Activity
D
-620/+160
-620/+118
Luc
Rev.
SV-40 poly-A signal
For. #1
For. #2
ATAAGCtgcAATAAA
E
Ω–323
Ω–264
Ω–246
Ω–160
Ω–118
myogenin
Lucif.
S17

Figure 4.5: Analysis of unique properties of the RGMc 5'UTR. A. The percentage of GC-content (*red*) or pyrimidine enrichment (*green*) over a sliding window of 25-nucleotides across the 5'UTR of mouse RGMc. In addition, the first 16/22 nucleotides (73%) are pyrimidines (not shown) suggesting a possible 'terminal oligo-pyrimidine' (5'TOP) tract; see Fig. 4.6 for details. **B.** The position of various features of the RGMc 5'UTR including the exons 1 and 2, the exon-2 spice variants, and the region for the ε-element (from +118 to +160). Sequences with an ATG (possible AUG start codons, five uORFs and pAUG at +324) are shown as *shaded arrowheads* positioned at their respective positions on the sequence ruler. Well-conserved and strong Kozak consensus sequences (GCC(**A/G**)CCAUG**G**, -3 and +4 shaded in *bold* are the most critical residuces) are noted with asterisk, '*', (and see Fig. 4.6 for sequence details across different species). **C.** Sequence alignment of the ε-element (+118/+160, boxed and highlighted in *gray*), with the conserved ATG/AUG tri-nucleotide sequnce that does not meet a canonical 'strong Kozak-consensus' highlighted in *red*. **D.** Sequence alignment presented in **C** depicted as a logos to highlight the most well-conserved nucleotides across 9 mammalian species (alignment performed by hand and created using the WebLogos server at UC Berkeley, Refs.[225, 226]).

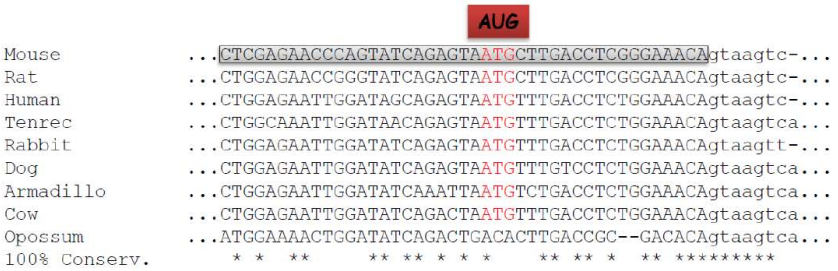
A


% C+G
% C+T
Window Size = 25
% Base Composition
100
80
60
40
20
0

B


exon 1
+160
exon 2
+323 (UTR)
+118
77-nt
18-nt
Position of ATGs
(and spacing)
1
100
200
300

C


AUG
Mouse ...CTCGAGAACCCAGTATCAGAGTAATGCTTGACCTCGGGAAACAgtaagtc-...
Rat ...CTGGAGAACCGGGTATCAGAGTAATGCTTGACCTCGGGAAACAgtaagtc-...
Human ...CTGGAGAATTGGATAGCAGAGTAATGTTTGACCTCTGGAAACAgtaagtc-...
Tenrec ...CTGGCAAATTGGATAACAGAGTAATGTTTGACCTCTGGAAACAgtaagtca...
Rabbit ...CTGGAGAATTGGATATCAGAGTAATGTTTGACCTCTGGAAACAgtaagtt-...
Dog ...CTGGAGAATTGGATATCAGAGTAATGTTTGTCCTCTGGAAACAgtaagtca...
Armadillo ...CTGGAGAATTGGATATCAAATTAATGTCTGACCTCTGGAAACAgtaagtca...
Cow ...CTGGAGAATTGGATATCAGACTAATGTTTGACCTCTGGAAACAgtaagtca...
Opossum ...ATGGAAAACTGGATATCAGACTGACACTTGACCGC--GACACAgtaagtca...
100% Conserv.

D

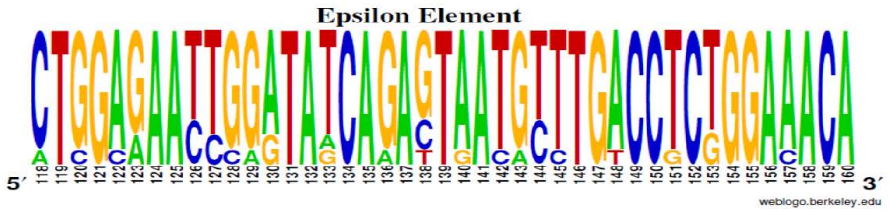

Epsilon Element
5′
3′
weblogo.berkeley.edu

Figure 4.6: Sequence alignment of the RGMc 5'UTR across multiple species. Alignment of genomic sequences from 10 mammalian species (exon 1 for elephant not present in current genomic sequence) in the 5' UTR of RGMc (exons in *upper case*), along with a portion of intron 1 (*in light gray lettering and lower case*). The mouse exon 2 highlighted as follows: the splice acceptor site that creates 174-nt exon 2 is in *black*, the additional 18-nt that creates the 192-nt variant is in *dark gray*, and the additional 77-nt found in the 251-nt variant is in *light gray*. ATG/AUG sequences are highlighted in *yellow*, and the defined protein coding sequence depicted in *blue*. The principal ATG/AUG start codon is noted in green. Additional details on the splice variants may be found in Fig. 3.3.


<---RGMc Promoter---> +1
Exon 1
Exon 1
Exon 1
<----------intron----------> 251-nt exon 2 variant
192-nt exon 2 variant
Exon 2 (174-nt variant)
pAUG
(pAUG = PRINCIPAL START CODON)
PROTEIN CODING SEQUENCE---->
<----------intron----------> ~1kb-->

Figure 4.7: RGMc promoter is largely inactive in three unique liver cell lines, but epsilon element dramatically increases reporter activity. The **B.** Results of luciferase assays in AML-12 (*white*), Hep3B (*gray hatched*), or HepG2 (*black*) cells transiently transfected with reporter genes containing different constructs of the mouse RGMc promoter and genomic locus. Cells were incubated in growth media (see '*experimental procedures*' for details) for 24 hr before analysis (mean of 2 independent experiments in duplicate for RGMc constructs and 5 for control plasmids, IGF-II promoter (p)3 - luciferase and TK-luciferase). Values for TK-luciferase were set to 100 relative luciferase activity units (average measurements were 8 $x10^3$ (AML-12 cells), 10 x 10^4 (HepG2 cells), or 3 $x10^4$ (Hep3B cells) light units/μg total protein/sec).

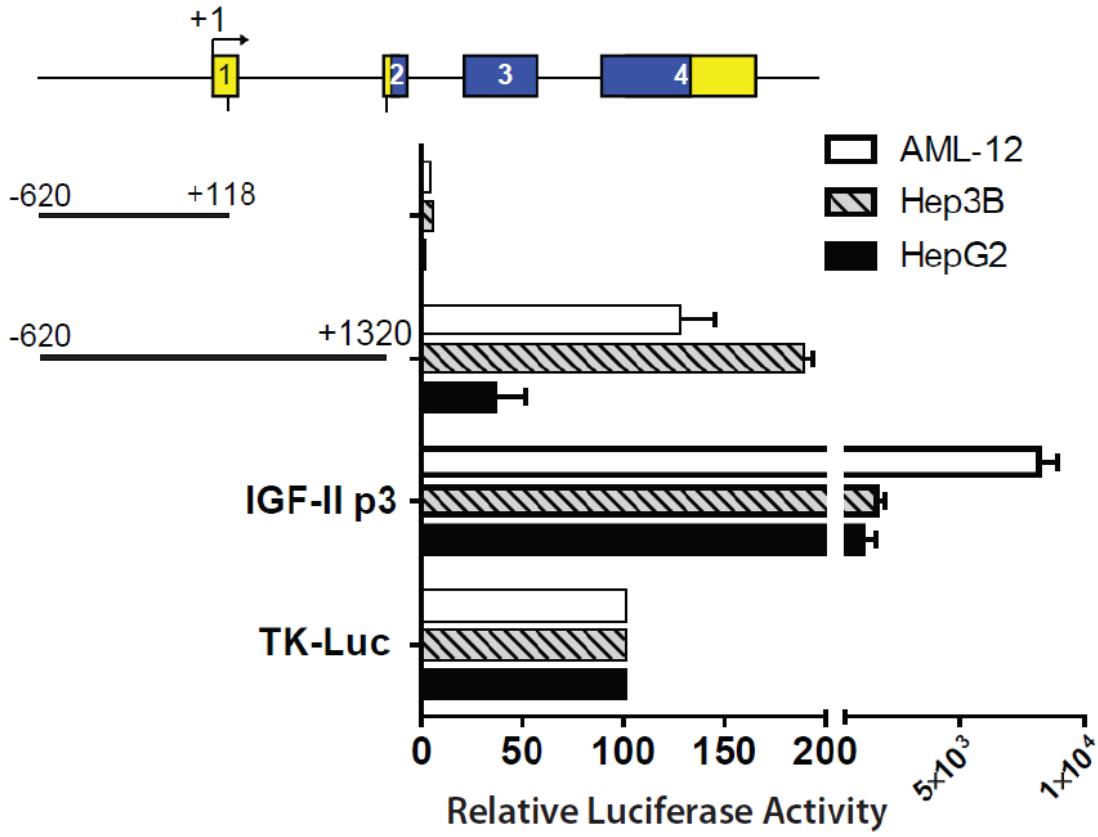

+1
1
2
3
4
AML-12
Hep3B
HepG2
-620
+118
-620
+1320
IGF-II p3
TK-Luc
0
50
100
150
200
5x10³
1x10⁴
Relative Luciferase Activity

Figure 4.8: Translational Control may be the major regulatory mechanism for RGMc expression in the Liver, and is not dependent upon iron levels. Hep3B cells were transfected with the 5'UTR "Omega" series (exon-2 variants) of RGMc reporter constucts with the length of the UTR indicated (see Fig. 4.4D for a diagram of the constructs). **A.** Luciferase Activity from RGMc Ω-series in Hep3Bs, with normalized levels as noted in Fig. 4.7. **B.** RT-PCR of transcripts from different RGMc Ω-series constructs in Hep3Bs. **C.** Iron loading: Ω-series with Hep3B cells under conditions of iron depletion (using 500 μM deferroxamine, DFO, in *white*), standard iron levels (*black*), and iron loading (with 500 μM ferric ammonium citrate, FAC, *gray hatched*). Results presented as the mean of two-independent experiments for DFO and FAC loaded cells, and n=5 experiments for controls, each in duplicate, with error bars representing the range. **D.** Western blot for the ferritin protein to show appropriate levels of iron depletion (with DFO), standard iron levels, and iron loading (with FAC). Relative molecular weight markers (in kDa) shown to the right.

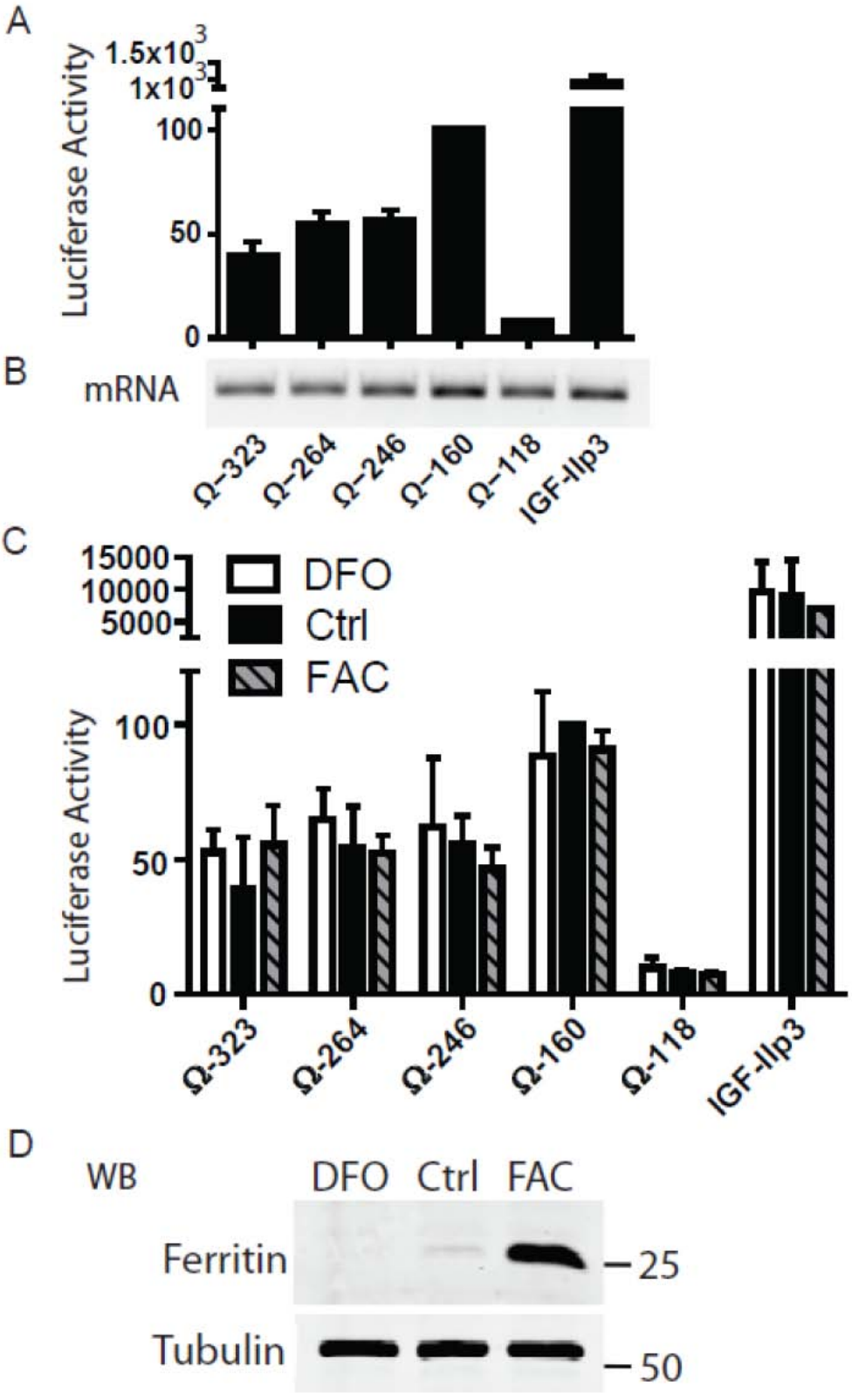
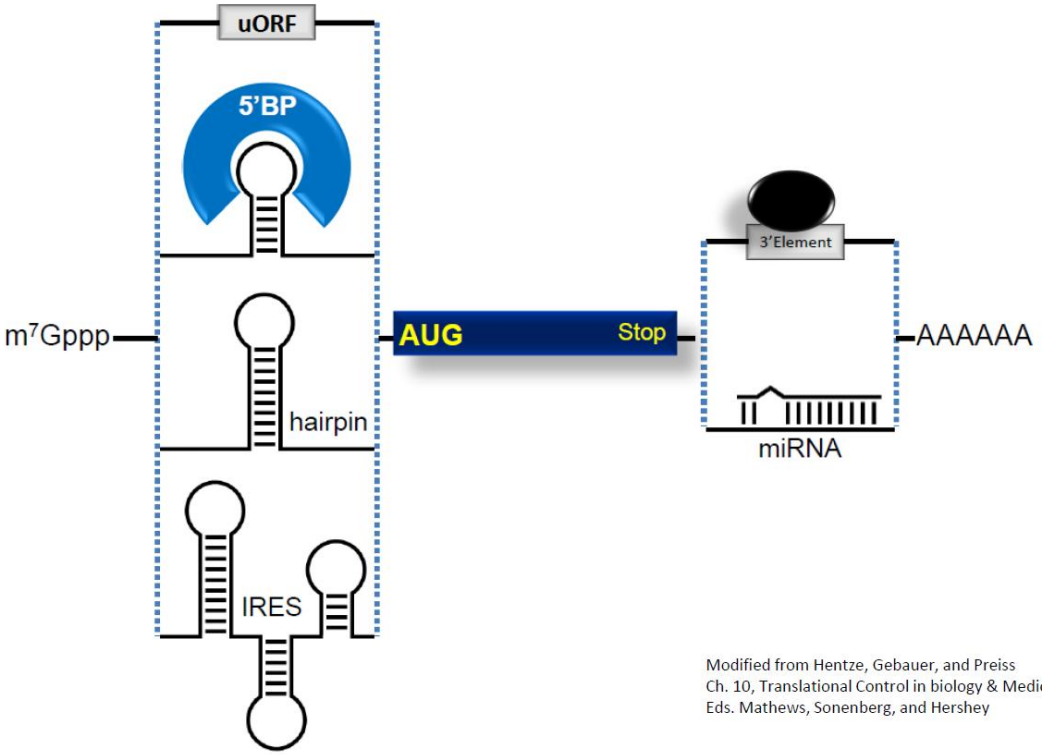

A
Luciferase Activity
1.5x10^3
1x10^3
100
50
0
B
mRNA
Ω-323
Ω-264
Ω-246
Ω-160
Ω-118
IGF-IIp3
C
15000
10000
5000
DFO
Ctrl
FAC
Luciferase Activity
100
50
0
Ω-323
Ω-264
Ω-246
Ω-160
Ω-118
IGF-IIp3
D
WB
DFO
Ctrl
FAC
Ferritin
25
Tubulin
50

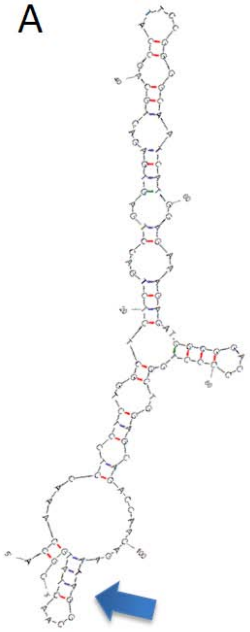
Figure 4.9: Mechanisms of the most common cis-acting elements in translation. The 5'-untranslated region (UTR) of an mRNA may contain an upstream open reading frame (uORF), a secondary structure that a binding protein (5'BP) may use to regulate translation or simply block a scanning ribosome (e.g., a hairpin), or form complex structures recognized as an internal ribosome entry site (IRES). The AUG start codon and stop codons flank the protein coding sequence (*dark blue*). Additional regulatory features in the 3'UTR including 3'elements with cognate binding proteins (*black oval*) or microRNA (miRNA) target sites may also regulate translation. Modified from Hentze, Gebauer, and Preiss, Ch. 10 of *Translational Control in Biology & Medicine*, Eds. Mathews, Sonenberg, and Hershey. Ref. [229].



Modified from Hentze, Gebauer, and Preiss
Ch. 10, Translational Control in biology & Medicine
Eds. Mathews, Sonenberg, and Hershey

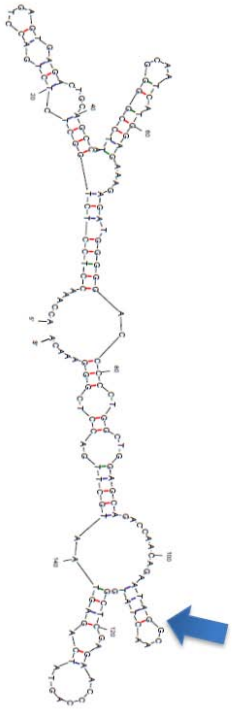
Figure 4.10: Computational prediction of secondary structure mRNA from the 5'UTR of RGMc. Representative results of the secondary structure of the 5'UTR from three different mRNA splice variants as predicted from mFold [257, 258]. C≡G bonds shown in *red,* and T=A bonds shown in *blue.* See Table 4.1 for additional details on the total number of structures and the average theoretical Gibbs Free Energy (ΔG^{o}) of folding (units kcal/mol) for the UTRs of RGMc. The *blue arrow* denotes hairpin structure at +104 to +115 that is present in the majority of predicted structures.

A



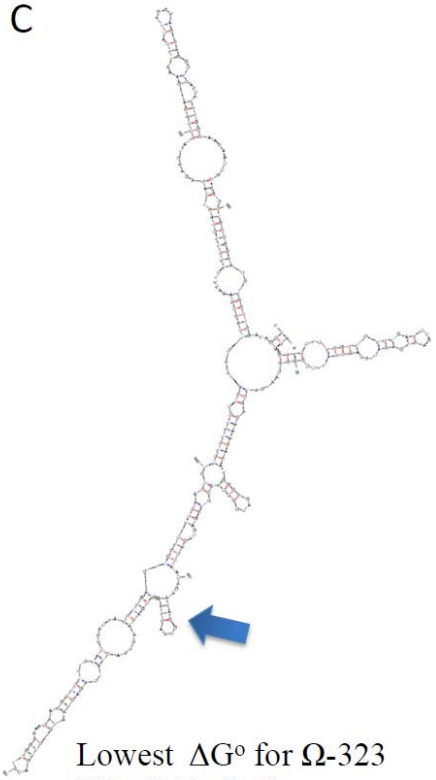
Ω-118
(ΔG°= -30.8 kcal/mol)

B



Ω-160
(ΔG°= -42.2 kcal/mol)

C



Lowest ΔG° for Ω-323
(ΔG°= -95.2 kcal/mol)

(This page was intentionally left blank)

Chapter 5

Summary and Future Directions

"To solve a basic problem in medicine, don't study it directly; rather, pursue a curiosity about nature and the rest will follow." –Roger Kornberg, 2008

"Somewhere, something incredible is waiting to be known." –Carl Sagan

5.1: Overview

The major findings and contributions of this dissertation are (i) defining the detailed gene structure of mouse RGMc, and using genomic sequence alignments to map the RGMc gene in other species, (ii) characterizing the promoter of RGMc in skeletal muscle, (iii) identifying a post-transcriptional regulatory element in the untranslated region of RGMc that appears to act as a positive translational regulator, and (iv) creating a phylogentic tree and *ab initio* protein model of the RGM family. In the text below, a synopsis of the major discoveries is outlined along with areas that I envision are the most important directions for future research within the field.

5.2: Summary of Chapter 2: The Repulsive Guidance Molecule Family

The work presented in chapter 2 analyzed sequencing and genomics data for the RGM family of genes, as well as summarized the current understanding of the published literature on the RGMs. As there had been no comprehensive assessment of the most fundamental aspects of the biology of the RGM family, including regulation of gene expression, control of protein biosynthesis, the relationship of protein structure to function, or mechanisms of action of each of the RGM proteins, the work in chapter 2 addressed the molecular biology and biochemistry of the RGM family, defined and critically evaluated what was known, and identified new areas for future investigation. A central theme throughout the chapter was comparative genomics from the chromosomal level, through the gene structure, to analysis of the protein sequence in order to make predictions about RGM family structure and function.

A phylogentic tree of the RGM family (Fig. 2.12) was developed through comparisons of the RGM family on the level of the chromosomal loci (Figs. 2.1, 2.6, an 2.8), the

structure of the genes (Figs. 2.2, 2.7, and 2.9), as well as analysis of the individual nucleotides on a "codon-optimized alignment" (see figure legend 2.12 for details). From this data, it was concluded that RGMa and RGMb shared a common ancestral gene more recently than RGMc. While there are features of the proteins that would suggest an alternative hypothesis (that RGMa and RGMc shared a common ancestral gene more recently, a view advocated by Camus and Lambert [35]), closer analysis of the data presented in this thesis and elsewhere [29, 152] suggest the relationship depicted in figure 2.12 is the most parsimonious explanation for the evolution of the RGM family. From a regulatory perspective, it is intriguing to speculate about the differences in structure and function of the promoters of the RGM family members. For example, what changes in the promoter allowed such unique tissue-restricted expression, with RGMa and RGMb being expressed in the nervous system with non-overlapping patterns, and RGMc being expressed exclusively in striated muscle and liver? Future work on the promoters of RGMa, RGMb, and species with a single RGM like *Ciona* will illuminate this interesting question.

On the level of the protein, it was shown that RGMa, RGMb, and RGMc, are products of distinct single-copy genes that arose early in vertebrate evolution, are ~ 40 - 50% identical to each other in primary amino acid sequence, and share similarities in predicted protein domains and overall structure. While this chapter focused primarily on previously published data and making inferences based on sequence homology, a recent publication by Nili, Shinde, and Rotwein demonstrated that the 50 and 40 kDa RGMc isoforms (see soluble forms in Figs. 2.3B and 2.5) can function as broad BMP antagonists [263], providing further insight into the possible function of the endogenous RGM proteins, in addition to being a valuable resource for fields that encomapsses BMP signaling, which range from development to cancer. The known disease-causing mutions of juvenile hemochomatosis (JH) were analyzed in comparison to the linear sequence

(Fig. 2.10 and Tabel 2.8), and coupled with the first *ab initio* protein model for any RGM gene. In sum, these data may provide critical insight into developing and testing new hypotheses about structure-function relationships in the RGM protein family.

5.3: Summary of Chapter 3: Structure of the RGMc gene and characterization of the RGMc promoter

Chapter 3 provided a look into the transcriptional regulatory mechanisms of RGMc, the first for any of the RGM family. The detailed gene structure of mouse RGMc was obtained via mapping of the transcription start site through 5'RACE and overlapping-primer RT-PCR experiments. I showed that RGMc is a 4-exon gene that undergoes alternative RNA splicing in exon 2 to yield three distinct mRNAs differing in the length of the 5' untranslated region (UTR). The possible functional consequences of these variable 5'UTRs was discussed in chapter 4 (see below). As this alternative splicing is present in all tissues known to express RGMc (Figs. 3.3 and 4.4A) and the splice acceptor site appears conserved across 10 mammalian vertebrates (Fig. 3.3B), it is likely that this variable 5'UTR is a property of all RGMc genes. As the 5'UTR of RGMc is longer than average (Fig. 3.4), the implications for additional mechanisms of regulation become apparent and are discussed in detail in chapter 4 (and see discussion below).

After demonstrating that the major regulatory mechanism for the appearance of mRNA is the induction of transcription (Figs. 3.5 and 3.6), I showed that RGMc is regulated by three defined regions in the proximal promoter (Fig. 3.10), that when collectively mutated, abrogate reporter activity (Fig. 3.13). Further analysis revealed that there are no other transcriptional enhancer, or repressor, elements located within a ~11.7 kb locus (Fig. 3.12), suggesting that the major transcriptional regulatory elements for RGMc

transcription in skeletal muscle are located within a well-conserved (Figs. 3.8 and 3.14) region of ~0.6 kb.

These three critical regions of the proximal RGMc gene promoter, comprising paired E-boxes, a putative Stat and/or Ets element, and a MEF2 site, are controlled by members of the bHLH and MEF2 family of transcription factors (Figs. 3.15 and 3.16). Future work will need to identify the factors necessary for the regulation of the β-element (summarized in figure 3.17). In addition, experiments to understand the transcriptional regulation in the liver will likely provide a fuller appreciation for evolution of the RGMc promoter, and how its expression is restricted to striated muscle and liver. A survey of published large-scale experiments presented in chapter 3 suggests that the orphan nuclear receptor HNF4α is localized to the locus [183], and functionally important [184] for RGMc expression. Future work will need to uncover these mechanisms as well as the understanding of the chromatin remodeling machinery associated with the regulation of RGMc, and the RGM family of genes.

Finally, using phylogenetic footprinting based on the functional data presented in the rest of the chapter, I suggest several hypotheses about the regulatory mechanisms that may control the expression of RGMc in other species, along with inferences that extend to the RGM family. The data presented in chapters 2 and 3 provide the research community a foundation with which to expand our knowledge about the regulation of the RGM family of genes, and understanding as to how genes evolve to have such unique tissue-restricted patterns of expression.

5.4: Summary of Chapter 4: Post-Transcriptional regulatory mechanisms of RGMc expression

The data in chapter 4 provide an additional regulatory mechanism for RGMc gene expression, and the intriguing possibility of a positive translational regulatory mechanism in the 5'UTR of the gene. I defined a novel post-transcriptional control region found within exon 1, called the ε-element, that increased reporter expression by 10-fold in muscle cells (Fig. 4.1) and ~40-fold in three liver cell lines (Fig 4.7). Three different hypotheses were presented (Fig 4.3) and tested in chapter 4. Since the levels of mRNA did not appreciably change with the addition of this region, the data presented in this chapter support the hypothesis that the ε-element is acting as a translational control element.

As noted in the introduction (chapter 1) of this thesis, regulation at the level of translation can provide numerous benefits over control at the level of transcription including rapidity and fine control of expression of the nascent protein [22]. One interesting possibility was that iron levels might change the levels of expression with each of the three different 5'UTRs found in RGMc (Figs. 3.2, 3.3, and 4.4). However, as shown in figure 4.8C, neither iron depletion nor iron loading causes a change in expression of RGMc. Thus, in combination with previously published work form others [238], it appears that RGMc levels may only be sensitive to iron levels at the protein level [67]. Nevertheless, there may be a subtle (i.e., ~2-fold) change in levels of expression with the longer sized 5'UTRs, but the functional consequence of these changes are currently unknown. Several possible reasons for this change were presented including oligopyrimidine binding proteins, secondary structures in the UTR, or structures permitting ribosome entry (e.g., IRES) or "shunting" (bypass of structures). Future work should begin with individual mutations of the uORFs (Fig. 4.5B), as well as an *in vitro* translation of the mRNA with and without the ε-element. A great deal of experimental work is simply waiting to uncover the mechanism by which the ε-element regulates RGMc expression.

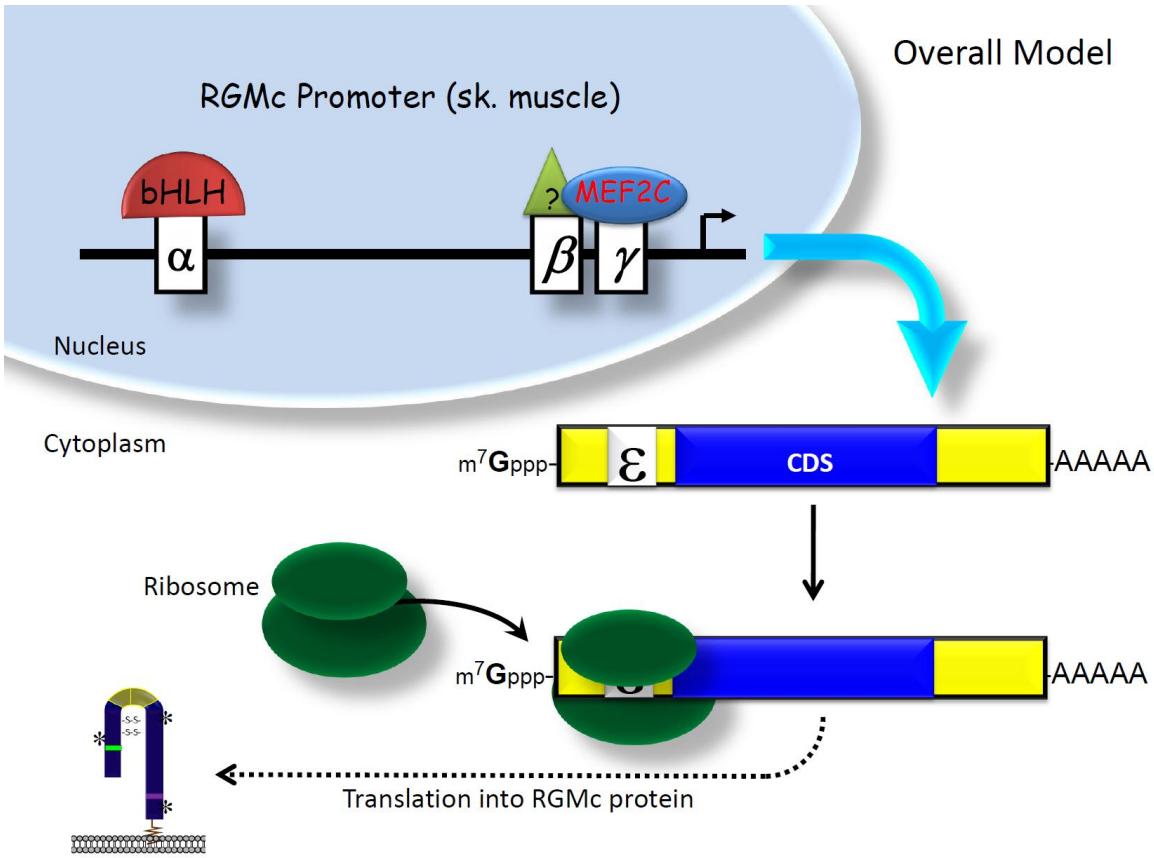
5.5: Concluding Statements

The results presented in this dissertation provide the foundation for understanding the regulatory mechanisms of RGMc expression and evolutionary origins of the family. These data also provide ample opportunity for follow-up experiments ranging from characterization of the RGMc promoter in other tissues and species, to promoter and post-transcriptional regulation in other RGM family members, to evolutionary analysis by comparing the structure and function of the RGM genes from animals with a single RGM (e.g., *Ciona intestinalis*) to those animals with multiple RGMs. From the data presented here, it appears likely that RGMc is a gene that arose early in vertebrate evolution, and may have a great deal to teach us concerning iron metabolism and tissue-specific gene regulation. In addition, the ancestral function of the RGM family remains unsolved, and the regulatory mechanisms of the other family members equally enigmatic. One of the primary perspectives presented in this dissertation may be summed up by a statement made by François Jacob in 1977:

> Evolution does not produce novelties from scratch. It works on what already exists, either transforming a system to give it new functions or combining several systems to produce a more elaborate one. This happened, for instance, during one of the main events of cellular evolution: namely, the passage from unicellular to multicellular forms. This was a particularly important transition because it carried an enormous potential for a specialization of the parts. Such a transition, which probably occurred several times, did not require the creation of new chemical species, for there are no major differences between molecular types of uni and multicellular organisms. It was mainly a reorganization of what already existed.
> -*Science* (1977), Ref. [26]

My hope is that the work presented here has contributed to the fundamental knowledge of RGMc and the RGM family of genes, as well as provide tools for additional hypotheses to be tested in understanding the molecular evolution of the appearance and regulation of gene families.

Figure 5.1: Model for the regulation of RGMc derived from data in this dissertation. The model for RGMc regulation begins with transcriptional control by three promoter elements (described in chapter 3). In muscle, the α-element is a set of paired E-boxes spaced ~70 base pairs apart and controlled by members of the basic helix-loop-helix (bHLH) family of transcription factors (TFs). In skeletal muscle, this is most likely myogenin. The β-element is a critical region for promoter activity and appears to be a Stat and/or Ets binding site, but the exact TF is unknown at this time. The γ-element is a MEF2 site bound by a member of the MEF2 family of TFs. Following the appearance of a nascent transcript, the ε-element in the 5'-untranslated region (UTR) appears to act as a positive regulator for translation. UTR in *yellow* and coding sequence (CDS) in *blue.*


Overall Model
RGMc Promoter (sk. muscle)
bHLH
α
?
MEF2C
β
γ
Nucleus
Cytoplasm
m7Gppp-
ε
CDS
-AAAAA
Ribosome
m7Gppp-
-AAAAA
Translation into RGMc protein

(This page was intentionally left blank)

REFERENCES

1. Niederkofler V, Salie R, Arber S: **Hemojuvelin is essential for dietary iron sensing, and its mutation leads to severe iron overload**. *J Clin Invest* 2005, **115**(8):2180-2186.

2. Huang FW, Pinkus JL, Pinkus GS, Fleming MD, Andrews NC: **A mouse model of juvenile hemochromatosis**. *J Clin Invest* 2005, **115**(8):2187-2191.

3. Papanikolaou G, Samuels ME, Ludwig EH, MacDonald ML, Franchini PL, Dube MP, Andres L, MacFarlane J, Sakellaropoulos N, Politou M *et al*: **Mutations in HFE2 cause iron overload in chromosome 1q-linked juvenile hemochromatosis**. *Nat Genet* 2004, **36**(1):77-82.

4. Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell JE, Jr. (eds.): **Molecular Cell Biology, 4/e**. New York, NY: W.H. Freeman and Company; 2000.

5. Carey MF, Peterson CL, Smale ST: **Transcriptional Regulation in Eukaryotes: Concepts, Strategies, and Techniques, 2/e**. Cold Spring Harbor, NY: Cold Spring Harbor Labs; 2009.

6. Keegan L, Gill G, Ptashne M: **Separation of DNA binding from the transcription-activating function of a eukaryotic regulatory protein**. *Science* 1986, **231**(4739):699-704.

7. Ptashne M: **How eukaryotic transcriptional activators work**. *Nature* 1988, **335**(6192):683-689.

8. Ptashne M: **A Genetic Switch, Third Edition, Phage Lambda Revisited**, 3 edn. Sloan-Kettering Memorial Cancer Center, New York CSHL Press; 2004.

9. Ptashne M: **Binding reactions: epigenetic switches, signal transduction and cancer**. *Current Biology* 2009, **19**(6):R234-R241.

10. Hochheimer A, Tjian R: **Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression**. *Genes & Development* 2003, **17**(11):1309-1320.

11. Li B, Carey M, Workman J: **The Role of Chromatin during Transcription** *Cell* 2007, **128**(5):707-719.

12. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (eds.): **Molecular Biology of the Cell**, 4 edn. New York, NY: Garland Science; 2002.

13. King MC, Wilson AC: **Evolution at two levels in humans and chimpanzees**. *Science* 1975, **188**(4184):107-116.

14. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA: **The evolution of transcriptional regulation in eukaryotes**. *Mol Biol Evol* 2003, **20**(9):1377-1419.

15. Lynch M: **The Origins of Eukaryotic Gene Structure**. *Mol Biol Evol* 2006, **23**(2):450-468.

16. Potthoff MJ, Olson EN: **MEF2: a central regulator of diverse developmental programs**. *Development* 2007, **134**(23):4131-4140.

17. Lynch M: **The Origins of Genome Architecture**. Sunderland, MA: Sinauer Assocs., Inc.; 2007.

18. Wray GA: **The evolutionary significance of cis-regulatory mutations**. *Nat Rev Genet* 2007, **8**(3):206-216.

19. Fischer D, Backendorf C: **Identification of Regulatory Elements by Gene Family Footprinting and In Vivo Analysis**. In: *Analytics of Protein–DNA Interactions.* 2007: 37-64.

20. Chen X, Tompa M: **Comparative assessment of methods for aligning multiple genome sequences**. *Nat Biotechnol* 2010, **Adv. Online publication**:doi:10.1038/nbt.1637.

21. Prakash A, Tompa M: **Discovery of regulatory elements in vertebrates through comparative genomics**. *Nat Biotechnol* 2005, **23**(10):1249-1256.

22. Mathews MB, Sonenberg N, Hershey JWB (eds.): **Translational Control in Biology and Medicine**. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2007.

23. Silvera D, Formenti SC, Schneider RJ: **Translational control in cancer**. *Nat Rev Cancer* 2010, **10**(4):254-266.

24. Pietrangelo A: **Hereditary hemochromatosis--a new look at an old disease**. *N Engl J Med* 2004, **350**(23):2383-2397.

25. Fleming RE, Bacon BR: **Orchestration of iron homeostasis**. *N Engl J Med* 2005, **352**(17):1741-1744.

26. Jacob F: **Evolution and tinkering**. *Science* 1977, **196**(4295):1161-1166.

27. Niederkofler V, Salie R, Sigrist M, Arber S: **Repulsive guidance molecule (RGM) gene function is required for neural tube closure but not retinal topography in the mouse visual system**. *J Neurosci* 2004, **24**(4):808-818.

28. Monnier PP, Sierra A, Macchi P, Deitinghoff L, Andersen JS, Mann M, Flad M, Hornberger MR, Stahl B, Bonhoeffer F *et al*: **RGM is a repulsive guidance molecule for retinal axons**. *Nature* 2002, **419**(6905):392-395.

29. Schmidtmer J, Engelkamp D: **Isolation and expression pattern of three mouse homologues of chick Rgm**. *Gene Expr Patterns* 2004, **4**(1):105-110.

30. Samad TA, Srinivasan A, Karchewski LA, Jeong SJ, Campagna JA, Ji RR, Fabrizio DA, Zhang Y, Lin HY, Bell E *et al*: **DRAGON: a member of the repulsive guidance molecule-related family of neuronal- and muscle-expressed membrane proteins is regulated by DRG11 and has neuronal adhesive properties**. *J Neurosci* 2004, **24**:2027 - 2036.

31. Kuninger D, Kuzmickas R, Peng B, Pintar JE, Rotwein P: **Gene discovery by microarray: identification of novel genes induced during growth factor-mediated muscle cell survival and differentiation**. *Genomics* 2004, **84**(5):876-889.

32. Matsunaga E, Tauszig-Delamasure S, Monnier PP, Mueller BK, Strittmatter SM, Mehlen P, Chedotal A: **RGM and its receptor neogenin regulate neuronal survival**. *Nat Cell Biol* 2004, **6**(8):749-755.

33. Oldekamp J, Kramer N, Alvarez-Bolado G, Skutella T: **Expression pattern of the repulsive guidance molecules RGM A, B and C during mouse development**. *Gene Expr Patterns* 2004, **4**(3):283-288.

34. Kuninger D, Kuns-Hashimoto R, Kuzmickas R, Rotwein P: **Complex biosynthesis of the muscle-enriched iron regulator RGMc**. *J Cell Sci* 2006, **119**(Pt 16):3273-3283.

35. Camus LM, Lambert LA: **Molecular evolution of hemojuvelin and the repulsive guidance molecule family**. *J Mol Evol* 2007, **65**(1):68-81.

36. Catchen JM, Conery JS, Postlethwait JH: **Inferring ancestral gene order**. *Methods Mol Biol* 2008, **452**:365-383.

37. Wheelan SJ, Church DM, Ostell JM: **Spidey: a tool for mRNA-to-genomic alignments**. *Genome Res* 2001, **11**(11):1952-1957.

38. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI**. *Nucleic Acids Res* 2005, **33**(Database issue):D54-58.

39. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**(5):1792-1797.

40. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends Genet* 2000, **16**(6):276-277.

41. Sprague J, Bayraktaroglu L, Clements D, Conlin T, Fashena D, Frazer K, Haendel M, Howe DG, Mani P, Ramachandran S *et al*: **The Zebrafish Information Network: the zebrafish model organism database**. *Nucleic Acids Res* 2006, **34**(Database issue):D581-585.

42. Consortium ICGS: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution**. *Nature* 2004, **432**(7018):695-716.

43. Matsunaga E, Nakamura H, Chedotal A: **Repulsive guidance molecule plays multiple roles in neuronal differentiation and axon guidance**. *J Neurosci* 2006, **26**(22):6082-6088.

44. Babitt JL, Zhang Y, Samad TA, Xia Y, Tang J, Campagna JA, Schneyer AL, Woolf CJ, Lin HY: **Repulsive guidance molecule (RGMa), a DRAGON homologue, is a bone morphogenetic protein co-receptor**. *J Biol Chem* 2005, **280**(33):29820-29827.

45. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC**. *Genome Res* 2002, **12**(6):996-1006.

46. Brinks H, Conrad S, Vogt J, Oldekamp J, Sierra A, Deitinghoff L, Bechmann I, Alvarez-Bolado G, Heimrich B, Monnier PP *et al*: **The repulsive guidance molecule RGMa is involved in the formation of afferent connections in the dentate gyrus**. *J Neurosci* 2004, **24**(15):3862-3869.

47. Doering TL, Schekman R: **GPI anchor attachment is required for Gas1p transport from the endoplasmic reticulum in COP II vesicles**. *EMBO J* 1996, **15**(1):182-191.

48. Ruoslahti E: **RGD and other recognition sequences for integrins**. *Annu Rev Cell Dev Biol* 1996, **12**:697-715.

49. Sadler JE: **Biochemistry and Genetics of von Willebrand Factor**. *Annual Review of Biochemistry* 1998, **67**(1):395-424.

50. Matsunaga E, Chédotal A: **Repulsive guidance molecule/neogenin: a novel ligand-receptor system playing multiple roles in neural development**. *Development Growth & Differentiation* 2004, **46**(6):481-486.

51. Stahl B, Muller B, von Boxberg Y, Cox EC, Bonhoeffer F: **Biochemical characterization of a putative axonal guidance molecule of the chick visual system**. *Neuron* 1990, **5**(5):735-743.

52. Hata K, Fujitani M, Yasuda Y, Doya H, Saito T, Yamagishi S, Mueller BK, Yamashita T: **RGMa inhibition promotes axonal growth and recovery after spinal cord injury**. *J Cell Biol* 2006, **173**(1):47-58.

53. Rajagopalan S, Deitinghoff L, Davis D, Conrad S, Skutella T, Chedotal A, Mueller BK, Strittmatter SM: **Neogenin mediates the action of repulsive guidance molecule**. *Nat Cell Biol* 2004, **6**(8):756-762.

54. Cirulli V, Yebra M: **Netrins: beyond the brain**. *Nat Rev Mol Cell Biol* 2007, **8**(4):296-306.

55. Hata K, Kaibuchi K, Inagaki S, Yamashita T: **Unc5B associates with LARG to mediate the action of repulsive guidance molecule**. *J Cell Biol* 2009, **184**(5):737-750.

56. Endo M, Yamashita T: **Inactivation of Ras by p120GAP via Focal Adhesion Kinase Dephosphorylation Mediates RGMa-Induced Growth Cone Collapse**. *J Neurosci* 2009, **29**(20):6649-6662.

57. Conrad S, Genth H, Hofmann F, Just I, Skutella T: **Neogenin-RGMa signaling at the growth cone is bone morphogenetic protein-independent and involves RhoA, ROCK, and PKC**. *J Biol Chem* 2007, **282**(22):16423-16433.

58. Schaffar G, Taniguchi J, Brodbeck T, Meyer AH, Schmidt M, Yamashita T, Mueller BK: **LIM-Only-Protein 4 (LMO4) interacts directly with the RGM A Receptor Neogenin**. *J Neurochem* 2008.

59. Xia Y, Yu PB, Sidis Y, Beppu H, Bloch KD, Schneyer AL, Lin HY: **Repulsive guidance molecule RGMa alters utilization of bone morphogenetic protein (BMP) type II receptors by BMP2 and BMP4**. *J Biol Chem* 2007, **282**(25):18129-18140.

60. Massague J: **TGF–β Signal Transduction**. *Annual Review of Biochemistry* 1998, **67**(1):753-791.

61. Ding YQ, Yin J, Xu HM, Jacquin MF, Chen ZF: **Formation of whisker-related principal sensory nucleus-based lemniscal pathway requires a paired homeodomain transcription factor, Drg11**. *J Neurosci* 2003, **23**(19):7246-7254.

62. Saito T, Greenwood A, Sun Q, Anderson DJ: **Identification by Differential RT-PCR of a Novel Paired Homeodomain Protein Specifically Expressed in Sensory Neurons and a Subset of Their CNS Targets** *Molecular and Cellular Neuroscience* 1995, **6**(3):280-292.

63. Schnichels S, Conrad S, Warstat K, Henke-Fahle S, Skutella T, Schraermeyer U, Julien S: **Gene expression of the repulsive guidance molecules/neogenin in the developing**

and mature mouse visual system: C57BL/6J vs. the glaucoma model DBA/2J. *Gene Expr Patterns* 2007, **8**(1):1-11.

64. Xia Y, Sidis Y, Mukherjee A, Samad TA, Brenner G, Woolf CJ, Lin HY, Schneyer A: **Localization and action of Dragon (repulsive guidance molecule b), a novel bone morphogenetic protein coreceptor, throughout the reproductive axis**. *Endocrinology* 2005, **146**(8):3614-3621.

65. Samad TA, Rebbapragada A, Bell E, Zhang Y, Sidis Y, Jeong S-J, Campagna JA, Perusini S, Fabrizio DA, Schneyer AL *et al*: **DRAGON, a Bone Morphogenetic Protein Co-receptor**. *J Biol Chem* 2005, **280**(14):14122-14129.

66. Andriopoulos Jr B, Corradini E, Xia Y, Faasse SA, Chen S, Grgurevic L, Knutson MD, Pietrangelo A, Vukicevic S, Lin HY *et al*: **BMP6 is a key endogenous regulator of hepcidin expression and iron metabolism**. *Nat Genet* 2009, **41**(4):482-487.

67. Kuninger D, Kuns-Hashimoto R, Nili M, Rotwein P: **Pro-protein convertases control the maturation and processing of the iron-regulatory protein, RGMc/hemojuvelin**. *BMC Biochemistry* 2008, **9**(1):9.

68. Silvestri L, Pagani A, Camaschella C: **Furin-mediated release of soluble hemojuvelin: a new link between hypoxia and iron homeostasis**. *Blood* 2008, **111**(2):924-931.

69. Lin L, Nemeth E, Goodnough JB, Thapa DR, Gabayan V, Ganz T: **Soluble hemojuvelin is released by proprotein convertase-mediated cleavage at a conserved polybasic RNRR site**. *Blood Cells Mol Dis* 2008, **40**(1):122-131.

70. Kuns-Hashimoto R, Kuninger D, Nili M, Rotwein P: **Selective Binding of RGMc/Hemojuvelin, a Key Protein in Systemic Iron Metabolism, to BMP-2 and Neogenin**. *Am J Physiol Cell Physiol* 2008.

71. Zhang AS, West AP, Jr., Wyman AE, Bjorkman PJ, Enns CA: **Interaction of hemojuvelin with neogenin results in iron accumulation in human embryonic kidney 293 cells**. *J Biol Chem* 2005, **280**(40):33885-33894.

72. Lin L, Goldberg YP, Ganz T: **Competitive regulation of hepcidin mRNA by soluble and cell-associated hemojuvelin**. *Blood* 2005, **106**:2884 - 2889.

73. Huang FW, Pinkus JL, Pinkus GS, Fleming MD, Andrews NC: **A mouse model of juvenile hemochromatosis**. *J Clin Invest* 2005, **115**:2187 - 2191.

74. Nemeth E, Tuttle MS, Powelson J, Vaughn MB, Donovan A, Ward DM, Ganz T, Kaplan J: **Hepcidin regulates cellular iron efflux by binding to ferroportin and inducing its internalization**. *Science* 2004, **306**:2090 - 2093.

75. Papanikolaou G, Tzilianos M, Christakis JI, Bogdanos D, Tsimirika K, MacFarlane J, Goldberg YP, Sakellaropoulos N, Ganz T, Nemeth E: **Hepcidin in iron overload disorders**. *Blood* 2005, **105**(10):4103-4105.

76. Nemeth E, Roetto A, Garozzo G, Ganz T, Camaschella C: **Hepcidin is decreased in TFR2 hemochromatosis**. *Blood* 2005, **105**(4):1803-1806.

77. Babitt JL, Huang FW, Wrighting DM, Xia Y, Sidis Y, Samad TA, Campagna JA, Chung RT, Schneyer AL, Woolf CJ *et al*: **Bone morphogenetic protein signaling by hemojuvelin regulates hepcidin expression**. *Nat Genet* 2006, **38**(5):531-539.

78. Babitt JL, Huang FW, Xia Y, Sidis Y, Andrews NC, Lin HY: **Modulation of bone morphogenetic protein signaling in vivo regulates systemic iron balance**. *J Clin Invest* 2007, **117**(7):1933-1939.

79. Xia Y, Babitt JL, Sidis Y, Chung RT, Lin HY: **Hemojuvelin regulates hepcidin expression via a selective subset of BMP ligands and receptors independently of neogenin**. *Blood* 2008, **111**(10):5195-5204.

80. Yang F, West AP, Jr., Allendorph GP, Choe S, Bjorkman PJ: **Neogenin interacts with hemojuvelin through its two membrane-proximal fibronectin type III domains**. *Biochemistry* 2008, **47**(14):4237-4245.

81. Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM *et al*: **The Draft Genome of Ciona intestinalis: Insights into Chordate and Vertebrate Origins**. *Science* 2002, **298**(5601):2157-2167.

82. Sea Urchin Genome Sequencing C, Sodergren E, Weinstock GM, Davidson EH, Cameron RA, Gibbs RA, Angerer RC, Angerer LM, Arnone MI, Burgess DR *et al*: **The**

Genome of the Sea Urchin Strongylocentrotus purpuratus. *Science* 2006, **314**(5801):941-952.

83. Consortium CeS: **Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology**. *Science* 1998, **282**(5396):2012-2018.

84. Eswar N, Eramian D, Webb B, Shen MY, Sali A: **Protein structure modeling with MODELLER**. *Methods Mol Biol* 2008, **426**:145-159.

85. Sali A, Potterton L, Yuan F, van Vlijmen H, Karplus M: **Evaluation of comparative protein modeling by MODELLER**. *Proteins* 1995, **23**(3):318-326.

86. Jones DT, Taylor WR, Thornton JM: **A new approach to protein fold recognition**. *Nature* 1992, **358**(6381):86-89.

87. Kim DE, Chivian D, Baker D: **Protein structure prediction and analysis using the Robetta server**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W526-531.

88. Simons KT, Kooperberg C, Huang E, Baker D: **Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions**. *J Mol Biol* 1997, **268**(1):209-225.

89. Bonneau R, Strauss CE, Rohl CA, Chivian D, Bradley P, Malmstrom L, Robertson T, Baker D: **De novo prediction of three-dimensional structures for major protein families**. *J Mol Biol* 2002, **322**(1):65-78.

90. Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CE, Baker D: **Rosetta in CASP4: progress in ab initio protein structure prediction**. *Proteins* 2001, **Suppl 5**:119-126.

91. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D: **Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins**. *Proteins* 1999, **34**(1):82-95.

92. Chivian D, Kim DE, Malmstrom L, Schonbrun J, Rohl CA, Baker D: **Prediction of CASP6 structures using automated Robetta protocols**. *Proteins* 2005, **61 Suppl 7**:157-166.

93. Chivian D, Kim DE, Malmstrom L, Bradley P, Robertson T, Murphy P, Strauss CE, Bonneau R, Rohl CA, Baker D: **Automated prediction of CASP-5 structures using the Robetta server**. *Proteins* 2003, **53 Suppl 6**:524-533.

94. Das R, Qian B, Raman S, Vernon R, Thompson J, Bradley P, Khare S, Tyka MD, Bhat D, Chivian D *et al*: **Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home**. *Proteins: Structure, Function, and Bioinformatics* 2007, **69**(S8):118-128.

95. Jorieux S, Fressinaud E, Goudemand J, Gaucher C, Meyer D, Mazurier C, INSERM: **Conformational changes in the D' domain of von Willebrand factor induced by CYS 25 and CYS 95 mutations lead to factor VIII binding defect and multimeric impairment (INSERM (Inserm Network on Molecular Abnormalities in von Willebrand Disease) group)**. *Blood* 2000, **95**(10):3139-3145.

96. Fukuda K, Doggett TA, Bankston LA, Cruz MA, Diacovo TG, Liddington RC: **Structural Basis of von Willebrand Factor Activation by the Snake Toxin Botrocetin** *Structure* 2002, **10**(7):943-950.

97. Downing AK, Knott V, Werner JM, Cardy CM, Campbell ID, Handford PA: **Solution Structure of a Pair of Calcium-Binding Epidermal Growth Factor-like Domains: Implications for the Marfan Syndrome and Other Genetic Disorders**. *Cell* 1996, **85**(4):597-605.

98. Herlyn H, Zischler H: **The molecular evolution of sperm zonadhesin**. *Int J Dev Biol* 2008, **52**(5-6):781-790.

99. Lidell ME, Johansson MEV, Hansson GC: **An Autocatalytic Cleavage in the C Terminus of the Human MUC2 Mucin Occurs at the Low pH of the Late Secretory Pathway**. *J Biol Chem* 2003, **278**(16):13944-13951.

100. Lanzara C, Roetto A, Daraio F, Rivard S, Ficarella R, Simard H, Cox TM, Cazzola M, Piperno A, Gimenez-Roqueplo AP *et al*: **Spectrum of hemojuvelin gene mutations in 1q-linked juvenile hemochromatosis**. *Blood* 2004, **103**(11):4317-4321.

101. Villard V, Kalyuzhniy O, Riccio O, Potekhin S, Melnik TN, Kajava AV, Ruegg C, Corradin G: **Synthetic RGD-containing alpha-helical coiled coil peptides promote integrin-dependent cell adhesion**. *J Pept Sci* 2006, **12**(3):206-212.

102. Dube DH, Prescher JA, Quang CN, Bertozzi CR: **Probing mucin-type O-linked glycosylation in living animals**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(13):4819-4824.

103. Huang FW, Rubio-Aliaga I, Kushner JP, Andrews NC, Fleming MD: **Identification of a novel mutation (C321X) in HJV**. *Blood* 2004, **104**(7):2176-2177.

104. Lee PL, Beutler E, Rao SV, Barton JC: **Genetic abnormalities and juvenile hemochromatosis: mutations of the HJV gene encoding hemojuvelin**. *Blood* 2004, **103**(12):4669-4671.

105. Gehrke SG, Pietrangelo A, Kascak M, Braner A, Eisold M, Kulaksiz H, Herrmann T, Hebling U, Bents K, Gugler R *et al*: **HJV gene mutations in European patients with juvenile hemochromatosis**. *Clin Genet* 2005, **67**(5):425-428.

106. Aguilar-Martinez P, Lok CY, Cunat S, Cadet E, Robson K, Rochette J: **Juvenile hemochromatosis caused by a novel combination of hemojuvelin G320V/R176C mutations in a 5-year old girl**. *Haematologica* 2007, **92**(3):421-422.

107. Koyama C, Hayashi H, Wakusawa S, Ueno T, Yano M, Katano Y, Goto H, Kidokoro R: **Three patients with middle-age-onset hemochromatosis caused by novel mutations in the hemojuvelin gene**. *J Hepatol* 2005, **43**(4):740-742.

108. Wallace DF, Dixon JL, Ramm GA, Anderson GJ, Powell LW, Subramaniam N: **Hemojuvelin (HJV)-associated hemochromatosis: analysis of HJV and HFE mutations and iron overload in three families**. *Haematologica* 2005, **90**(2):254-255.

109. Daraio F, Ryan E, Gleeson F, Roetto A, Crowe J, Camaschella C: **Juvenile hemochromatosis due to G320V/Q116X compound heterozygosity of hemojuvelin in an Irish patient**. *Blood Cells Mol Dis* 2005, **35**(2):174-176.

110. Lee PL, Barton JC, Brandhagen D, Beutler E: **Hemojuvelin (HJV) mutations in persons of European, African-American and Asian ancestry with adult onset haemochromatosis**. *Br J Haematol* 2004, **127**(2):224-229.

111. Lee P, Promrat K, Mallette C, Flynn M, Beutler E: **A juvenile hemochromatosis patient homozygous for a novel deletion of cDNA nucleotide 81 of hemojuvelin**. *Acta Haematol* 2006, **115**(1-2):123-127.

112. Murugan RC, Lee PL, Kalavar MR, Barton JC: **Early age-of-onset iron overload and homozygosity for the novel hemojuvelin mutation HJV R54X (exon 3; c.160A-->T) in an African American male of West Indies descent**. *Clin Genet* 2008, **74**(1):88-92.

113. Janosi A, Andrikovics H, Vas K, Bors A, Hubay M, Sapi Z, Tordai A: **Homozygosity for a novel nonsense mutation (G66X) of the HJV gene causes severe juvenile hemochromatosis with fatal cardiomyopathy**. *Blood* 2005, **105**(1):432.

114. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucl Acids Res* 1994, **22**(22):4673-4680.

115. Suyama M, Torrents D, Bork P: **PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W609-612.

116. Guindon S, Gascuel O: **A simple, fast and accurate algorithm to estimate larges phylogenies by maximum likelihood**. *Systematic Biol* 2003, **52**:696-704.

117. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M *et al*: **Phylogeny.fr: robust phylogenetic analysis for the non-specialist**. *Nucleic Acids Res* 2008, **36**(Web Server issue):W465-469.

118. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models**. *Bioinformatics* 2003, **19**(12):1572-1574.

119. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees**. *Bioinformatics* 2001, **17**(8):754-755.

120. Stern A, Doron-Faigenboim A, Erez E, Martz E, Bacharach E, Pupko T: **Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W506-511.

121. Chevenet F, Brun C, Banuls A-L, Jacq B, Christen R: **TreeDyn: towards dynamic graphics and annotations for analyses of trees**. *BMC Bioinformatics* 2006, **7**(1):439.

122. Das R, Baker D: **Macromolecular Modeling with Rosetta**. *Annual Review of Biochemistry* 2008, **77**(1):363.

123. MacKerel Jr AD, Brooks Iii CL, Nilsson L, Roux B, Won Y, Karplus M. In: *CHARMM: The Energy Function and Its Parameterization with an Overview of the Program.* vol. 1: John Wiley & Sons: Chichester; 1998: 271-277.

124. Brooks BR, Bruccoleri RE, Olafson DJ, States DJ, Swaminathan S, Karplus M: **CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations**. *Journal of Computational Chemistry* 1983, **4**:187-217.

125. Subbian E, Yabuta Y, Shinde U: **Positive selection dictates the choice between kinetic and thermodynamic protein folding and stability in subtilases**. *Biochemistry* 2004, **43**(45):14348-14360.

126. Hentze MW, Muckenthaler MU, Andrews NC: **Balancing acts: molecular control of mammalian iron metabolism**. *Cell* 2004, **117**:285 - 297.

127. Pugh RA, Honda M, Leesley H, Thomas A, Lin Y, Nilges MJ, Cann IKO, Spies M: **The Iron-containing Domain Is Essential in Rad3 Helicases for Coupling of ATP Hydrolysis to DNA Translocation and for Targeting the Helicase to the Single-stranded DNA-Double-stranded DNA Junction**. *J Biol Chem* 2008, **283**(3):1732-1743.

128. Andrews NC, Schmidt PJ: **Iron Homeostasis**. *Annual Review of Physiology* 2007, **69**(1):69-85.

129. Zhang D-L, Hughes RM, Ollivierre-Wilson H, Ghosh MC, Rouault TA: **A Ferroportin Transcript that Lacks an Iron-Responsive Element Enables Duodenal and Erythroid Precursor Cells to Evade Translational Repression**. *Cell Metabolism* 2009, **9**(5):461-473.

130. Pietrangelo A, Caleffi A, Henrion J, Ferrara F, Corradini E, Kulaksiz H, Stremmel W, Andreone P, Garuti C: **Juvenile hemochromatosis associated with pathogenic mutations of adult hemochromatosis genes**. *Gastroenterology* 2005, **128**(2):470-479.

131. Wilson EM, Hsieh MM, Rotwein P: **Autocrine growth factor signaling by insulin-like growth factor-II mediates MyoD-stimulated myocyte maturation**. *J Biol Chem* 2003, **278**(42):41109-41113.

132. Grompe M, Jones SN, Loulseged H, Caskey CT: **Retroviral-mediated gene transfer of human ornithine transcarbamylase into primary hepatocytes of spf and spf-ash mice**. *Hum Gene Ther* 1992, **3**(1):35-44.

133. Overturf K, Al-Dhalimy M, Tanguay R, Brantly M, Ou CN, Finegold M, Grompe M: **Hepatocytes corrected by gene therapy are selected in vivo in a murine model of hereditary tyrosinaemia type I**. *Nat Genet* 1996, **12**(3):266-273.

134. Duncan AW, Hickey RD, Paulk NK, Culberson AJ, Olson SB, Finegold MJ, Grompe M: **Ploidy reductions in murine fusion-derived hepatocytes**. *PLoS Genet* 2009, **5**(2):e1000385.

135. Woelfle J, Billiard J, Rotwein P: **Acute Control of Insulin-like Growth Factor-I Gene Transcription by Growth Hormone through Stat5b**. *J Biol Chem* 2003, **278**(25):22696-22702.

136. Wilson EM, Rotwein P: **Control of MyoD function during initiation of muscle differentiation by an autocrine signaling pathway activated by insulin-like growth factor-II**. *J Biol Chem* 2006, **281**(40):29962-29971.

137. Kou K, Rotwein P: **Transcriptional activation of the insulin-like growth factor-II gene during myoblast differentiation**. *Mol Endocrinol* 1993, **7**(2):291-302.

138. Molkentin JD, Black BL, Martin JF, Olson EN: **Mutational analysis of the DNA binding, dimerization, and transcriptional activation domains of MEF2C**. *Mol Cell Biol* 1996, **16**(6):2627-2636.

139. Molkentin JD, Firulli AB, Black BL, Martin JF, Hustad CM, Copeland N, Jenkins N, Lyons G, Olson EN: **MEF2B is a potent transactivator expressed in early myogenic lineages**. *Mol Cell Biol* 1996, **16**(7):3814-3824.

140. Sambrook J, Russell DW: **Rapid amplification of 5′ cDNA ends (Protocol 9, 8.54–8.60)**, vol. Ch. 8. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press; 2001.

141. Sambrook J, Russell DW: **Rapid amplification of 5[prime] complementary DNA ends (5[prime] RACE)**. *Nat Meth* 2005, **2**(8):629-630.

142. Chia DJ, Ono M, Woelfle J, Schlesinger-Massart M, Jiang H, Rotwein P: **Characterization of distinct Stat5b binding sites that mediate growth hormone-stimulated IGF-I gene transcription**. *J Biol Chem* 2006, **281**(6):3190-3197.

143. Gronowski AM, Rotwein P: **Rapid changes in nuclear protein tyrosine phosphorylation after growth hormone treatment in vivo. Identification of phosphorylated mitogen-activated protein kinase and STAT91**. *J Biol Chem* 1994, **269**(11):7874-7878.

144. Liu N, Williams AH, Kim Y, McAnally J, Bezprozvannaya S, Sutherland LB, Richardson JA, Bassel-Duby R, Olson EN: **An intragenic MEF2-dependent enhancer directs muscle-specific expression of microRNAs 1 and 133**. *Proceedings of the National Academy of Sciences* 2007, **104**(52):20844-20849.

145. Ovcharenko I, Nobrega MA, Loots GG, Stubbs L: **ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W280-286.

146. Loots GG, Ovcharenko I: **Dcode.org anthology of comparative genomic tools**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W56-64.

147. Loots GG, Ovcharenko I: **rVISTA 2.0: evolutionary analysis of transcription factor binding sites**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W217-221.

148. Knuppel R, Dietze P, Lehnberg W, Frech K, Wingender E: **TRANSFAC retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins**. *J Comput Biol* 1994, **1**(3):191-198.

149. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank**. *Nucleic Acids Res* 2010, **38**(Database issue):D46-51.

150. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A: **JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles**. *Nucl Acids Res* 2009:gkp950.

151. Larsen F, Gundersen G, Lopez R, Prydz H: **CpG islands as gene markers in the human genome**. *Genomics* 1992, **13**(4):1095-1107.

152. Severyn CJ, Shinde U, Rotwein P: **Molecular biology, genetics and biochemistry of the repulsive guidance molecule family**. *Biochem J* 2009, **422**(3):393-403.

153. Bucher P: **Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences**. *Journal of Molecular Biology* 1990, **212**(4):563-578.

154. Kozak M: **An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs**. *Nucl Acids Res* 1987, **15**(20):8125-8148.

155. Pesole G, Mignone F, Gissi C, Grillo G, Licciulli F, Liuni S: **Structural and functional features of eukaryotic mRNA untranslated regions**. *Gene* 2001, **276**(1-2):73-81.

156. Meijer HA, Thomas AAM: **Control of eukaryotic protein synthesis by upstream open reading frames in the 5'-untranslated region of an mRNA**. *Biochem J* 2002, **367**(1):1-11.

157. Liu S, Zhang C, Zhou Y: **Uneven size distribution of mammalian genes in the number of tissues expressed and in the number of co-expressed genes**. *Hum Mol Genet* 2006, **15**(8):1313-1318.

158. Kozak M: **Features in the 5' non-coding sequences of rabbit alpha and beta-globin mRNAs that affect translational efficiency**. *J Mol Biol* 1994, **235**(1):95-110.

159. Tureckova J, Wilson EM, Cappalonga JL, Rotwein P: **Insulin-like growth factor-mediated muscle differentiation: collaboration between phosphatidylinositol 3-kinase-Akt-signaling pathways and myogenin**. *J Biol Chem* 2001, **276**(42):39264-39270.

160. Yaffe D, Saxel O: **Serial passaging and differentiation of myogenic cells isolated from dystrophic mouse muscle**. *Nature* 1977, **270**(5639):725-727.

161. Yaffe D, Saxel O: **A myogenic cell line with altered serum requirements for differentiation**. *Differentiation* 1977, **7**(3):159-166.

162. Hasty P, Bradley A, Morris JH, Edmondson DG, Venuti JM, Olson EN, Klein WH: **Muscle deficiency and neonatal death in mice with a targeted mutation in the myogenin gene**. *Nature* 1993, **364**(6437):501-506.

163. Nabeshima Y, Hanaoka K, Hayasaka M, Esuml E, Li S, Nonaka I, Nabeshima Y-i: **Myogenin gene disruption results in perinatal lethality because of severe muscle defect**. *Nature* 1993, **364**(6437):532-535.

164. Yamaguchi Y, Wada T, Handa H: **Interplay between positive and negative elongation factors: drawing a new view of DRB**. *Genes Cells* 1998, **3**(1):9-15.

165. Levy DE, Darnell JE, Jr.: **Stats: transcriptional control and biological impact**. *Nat Rev Mol Cell Biol* 2002, **3**(9):651-662.

166. Leonard WJ, O'Shea JJ: **JAKS AND STATS: Biological Implications***. *Annual Review of Immunology* 1998, **16**(1):293-322.

167. Horvath CM, Wen Z, Darnell JE, Jr.: **A STAT protein domain that determines DNA sequence recognition suggests a novel DNA-binding domain**. *Genes Dev* 1995, **9**(8):984-994.

168. Shore P, Whitmarsh AJ, Bhaskaran R, Davis RJ, Waltho JP, Sharrocks AD: **Determinants of DNA-binding specificity of ETS-domain transcription factors**. *Mol Cell Biol* 1996, **16**(7):3338-3349.

169. Sharrocks AD, Brown AL, Ling Y, Yates PR: **The ETS-domain transcription factor family**. *The International Journal of Biochemistry & Cell Biology* 1997, **29**(12):1371-1387.

170. Gutierrez-Hartmann A, Duval DL, Bradford AP: **ETS transcription factors in endocrine systems**. *Trends in Endocrinology & Metabolism* 1997, **18**(4):150-158.

171. Gossett LA, Kelvin DJ, Sternberg EA, Olson EN: **A new myocyte-specific enhancer-binding factor that recognizes a conserved element associated with multiple muscle-specific genes**. *Mol Cell Biol* 1989, **9**(11):5022-5033.

172. Sternberg EA, Spizz G, Perry WM, Vizard D, Weil T, Olson EN: **Identification of upstream and intragenic regulatory elements that confer cell-type-restricted and differentiation-specific expression on the muscle creatine kinase gene**. *Mol Cell Biol* 1988, **8**(7):2896-2909.

173. Trask RV, Strauss AW, Billadello JJ: **Developmental regulation and tissue-specific expression of the human muscle creatine kinase gene**. *Journal of Biological Chemistry* 1988, **263**(32):17142-17149.

174. Cserjesi P, Lilly B, Hinkley C, Perry M, Olson EN: **Homeodomain protein MHox and MADS protein myocyte enhancer-binding factor-2 converge on a common element in the muscle creatine kinase enhancer**. *Journal of Biological Chemistry* 1994, **269**(24):16740-16745.

175. Rudnicki MA, Schnegelsberg PN, Stead RH, Braun T, Arnold HH, Jaenisch R: **MyoD or Myf-5 is required for the formation of skeletal muscle**. *Cell* 1993, **75**(7):1351-1359.

176. Rudnicki MA, Braun T, Hinuma S, Jaenisch R: **Inactivation of MyoD in mice leads to up-regulation of the myogenic HLH gene Myf-5 and results in apparently normal muscle development**. *Cell* 1992, **71**(3):383-390.

177. Tapscott SJ: **The circuitry of a master switch: Myod and the regulation of skeletal muscle gene transcription**. *Development* 2005, **132**(12):2685-2695.

178. Lluis F, Perdiguero E, Nebreda AR, Munoz-Canoves P: **Regulation of skeletal muscle gene expression by p38 MAP kinases**. *Trends in Cell Biology* 2006, **16**(1):36-44.

179. Montgomery RL, Davis CA, Potthoff MJ, Haberland M, Fielitz J, Qi X, Hill JA, Richardson JA, Olson EN: **Histone deacetylases 1 and 2 redundantly regulate cardiac morphogenesis, growth, and contractility**. *Genes Dev* 2007, **21**(14):1790-1802.

180. Otu HH, Naxerova K, Ho K, Can H, Nesbitt N, Libermann TA, Karp SJ: **Restoration of Liver Mass after Injury Requires Proliferative and Not Embryonic Transcriptional Patterns**. *J Biol Chem* 2007, **282**(15):11197-11204.

181. Dhe-Paganon S, Duda K, Iwamoto M, Chi Y-I, Shoelson SE: **Crystal Structure of the HNF4Î± Ligand Binding Domain in Complex with Endogenous Fatty Acid Ligand**. *Journal of Biological Chemistry* 2002, **277**(41):37973-37976.

182. Takahashi H, Martin-Brown S, Washburn MP, Florens L, Conaway JW, Conaway RC: **Proteomics reveals a physical and functional link between hepatocyte nuclear factor 4alpha and transcription factor IID**. *J Biol Chem* 2009, **284**(47):32405-32412.

183. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S *et al*: **Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding**. *Science* 2010, **328**(5981):1036-1040.

184. Battle MA, Konopka G, Parviz F, Gaggl AL, Yang C, Sladek FM, Duncan SA: **Hepatocyte nuclear factor 4{alpha} orchestrates expression of cell adhesion proteins during the epithelial transformation of the developing liver**. *Proceedings of the National Academy of Sciences* 2006, **103**(22):8419-8424.

185. Clayton DF, Darnell JE, Jr.: **Changes in liver-specific compared to common gene transcription during primary culture of mouse hepatocytes**. *Mol Cell Biol* 1983, **3**(9):1552-1561.

186. Early P, Rogers J, Davis M, Calame K, Bond M, Wall R, Hood L: **Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways**. *Cell* 1980, **20**(2):313-319.

187. Nilsen TW, Graveley BR: **Expansion of the eukaryotic proteome by alternative splicing**. *Nature* 2010, **463**(7280):457-463.

188. Padgett RA, Grabowski PJ, Konarska MM, Seiler S, Sharp PA: **Splicing of Messenger RNA Precursors**. *Annual Review of Biochemistry* 1986, **55**(1):1119-1150.

189. Black DL: **MECHANISMS OF ALTERNATIVE PRE-MESSENGER RNA SPLICING**. *Annual Review of Biochemistry* 2003, **72**(1):291-336.

190. Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE, Zipursky SL: **Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity**. *Cell* 2000, **101**(6):671-684.

191. Kozak M: **Regulation of Translation in Eukaryotic Systems**. *Annual Review of Cell Biology* 1992, **8**(1):197-225.

192. Molkentin JD, Black BL, Martin JF, Olson EN: **Cooperative activation of muscle gene expression by MEF2 and myogenic bHLH proteins**. *Cell* 1995, **83**(7):1125-1136.

193. Dodou E, Xu SM, Black BL: **mef2c is activated directly by myogenic basic helix-loop-helix proteins during skeletal muscle development in vivo**. *Mech Dev* 2003, **120**(9):1021-1032.

194. Wang K, Wang C, Xiao F, Wang H, Wu Z: **JAK2/STAT2/STAT3 are required for myogenic differentiation**. *J Biol Chem* 2008, **283**(49):34029-34036.

195. Sun L, Ma K, Wang H, Xiao F, Gao Y, Zhang W, Wang K, Gao X, Ip N, Wu Z: **JAK1-STAT1-STAT3, a key pathway promoting proliferation and preventing premature differentiation of myoblasts**. *J Cell Biol* 2007, **179**(1):129-138.

196. De Val S, Anderson JP, Heidt AB, Khiem D, Xu S-M, Black BL: **Mef2c is activated directly by Ets transcription factors through an evolutionarily conserved endothelial cell-specific enhancer**. *Developmental Biology* 2004, **275**(2):424-434.

197. Verzi MP, Anderson JP, Dodou E, Kelly KK, Greene SB, North BJ, Cripps RM, Black BL: **N-twist, an evolutionarily conserved bHLH protein expressed in the developing CNS, functions as a transcriptional inhibitor**. *Dev Biol* 2002, **249**(1):174-190.

198. Lomvardas S, Thanos D: **Opening chromatin**. *Mol Cell* 2002, **9**(2):209-211.

199. Goodman RH, Smolik S: **CBP/p300 in cell growth, transformation, and development**. *Genes Dev* 2000, **14**(13):1553-1577.

200. Yao TP, Oh SP, Fuchs M, Zhou ND, Ch'ng LE, Newsome D, Bronson RT, Li E, Livingston DM, Eckner R: **Gene dosage-dependent embryonic development and proliferation defects in mice lacking the transcriptional integrator p300**. *Cell* 1998, **93**(3):361-372.

201. Puri PL, Sartorelli V, Yang XJ, Hamamori Y, Ogryzko VV, Howard BH, Kedes L, Wang JY, Graessmann A, Nakatani Y *et al*: **Differential roles of p300 and PCAF acetyltransferases in muscle differentiation**. *Mol Cell* 1997, **1**(1):35-45.

202. Polesskaya A, Duquet A, Naguibneva I, Weise C, Vervisch A, Bengal E, Hucho F, Robin P, Harel-Bellan A: **CREB-binding protein/p300 activates MyoD by acetylation**. *J Biol Chem* 2000, **275**(44):34359-34364.

203. Simone C, Forcales SV, Hill DA, Imbalzano AN, Latella L, Puri PL: **p38 pathway targets SWI-SNF chromatin-remodeling complex to muscle-specific loci**. *Nat Genet* 2004, **36**(7):738-743.

204. Ohkawa Y, Yoshimura S, Higashi C, Marfella CGA, Dacwag CS, Tachibana T, Imbalzano AN: **Myogenin and the SWI/SNF ATPase Brg1 Maintain Myogenic Gene Expression at Different Stages of Skeletal Myogenesis**. *J Biol Chem* 2007, **282**(9):6564-6570.

205. de la Serna IL, Ohkawa Y, Berkes CA, Bergstrom DA, Dacwag CS, Tapscott SJ, Imbalzano AN: **MyoD Targets Chromatin Remodeling Complexes to the Myogenin Locus Prior to Forming a Stable DNA-Bound Complex**. *Mol Cell Biol* 2005, **25**(10):3997-4009.

206. Soutoglou E, Katrakili N, Talianidis I: **Acetylation Regulates Transcription Factor Activity at Multiple Levels** *Molecular Cell* 2000, **5**(4):745-751.

207. Soutoglou E, Katrakili N, Talianidis I: **Acetylation regulates transcription factor activity at multiple levels**. *Mol Cell* 2000, **5**(4):745-751.

208. Torres-Padilla ME, Sladek FM, Weiss MC: **Developmentally Regulated N-terminal Variants of the Nuclear Receptor Hepatocyte Nuclear Factor 4alpha Mediate Multiple Interactions through Coactivator and Corepressor-Histone Deacetylase Complexes**. *J Biol Chem* 2002, **277**(47):44677-44687.

209. Gray SG, Ekstrom TJ: **The human histone deacetylase family**. *Exp Cell Res* 2001, **262**(2):75-83.

210. Lu J, McKinsey TA, Zhang C-L, Olson EN: **Regulation of Skeletal Myogenesis by Association of the MEF2 Transcription Factor with Class II Histone Deacetylases**. *Molecular Cell* 2000, **6**:233-244.

211. Mal AK: **Histone methyltransferase Suv39h1 represses MyoD-stimulated myogenic differentiation**. *EMBO J* 2006, **25**(14):3323-3334.

212. Caretti G, Di Padova M, Micales B, Lyons GE, Sartorelli V: **The Polycomb Ezh2 methyltransferase regulates muscle gene expression and skeletal muscle differentiation**. *Genes Dev* 2004, **18**(21):2627-2638.

213. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D: **Ultraconserved elements in the human genome**. *Science* 2004, **304**(5675):1321-1325.

214. Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio LA, Rubin EM: **Deletion of Ultraconserved Elements Yields Viable Mice**. *PLoS Biol* 2007, **5**(9):e234.

215. Canestro C, Yokoi H, Postlethwait JH: **Evolutionary developmental biology and genomics**. *Nat Rev Genet* 2007, **8**(12):932-942.

216. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D: **Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes**. *Proceedings of the National Academy of Sciences* 2003, **100**(20):11484-11489.

217. Dean AM, Thornton JW: **Mechanistic approaches to the study of evolution: the functional synthesis**. *Nat Rev Genet* 2007, **8**(9):675-688.

218. Ortlund EA, Bridgham JT, Redinbo MR, Thornton JW: **Crystal Structure of an Ancient Protein: Evolution by Conformational Epistasis**. *Science* 2007, **317**(5844):1544-1548.

219. Amoutzias GD, Robertson DL, Van de Peer Y, Oliver SG: **Choose your partners: dimerization in eukaryotic transcription factors**. *Trends Biochem Sci* 2008, **33**(5):220-229.

220. Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E: **Tissue-specific transcriptional regulation has diverged significantly between human and mouse**. *Nat Genet* 2007, **39**(6):730-732.

221. Hoffman MM, Birney E: **An effective model for natural selection in promoters**. *Genome Research* 2010, **20**(5):685-692.

222. Wasserman WW, Sandelin A: **Applied bioinformatics for the identification of regulatory elements**. *Nat Rev Genet* 2004, **5**(4):276-287.

223. Hedges SB, Blair JE, Venturi ML, Shoe JL: **A molecular timescale of eukaryote evolution and the rise of complex multicellular life**. *BMC Evol Biol* 2004, **4**:2.

224. Engström PG, Ho Sui SJ, Drivenes Ø, Becker TS, Lenhard B: **Genomic regulatory blocks underlie extensive microsynteny conservation in insects**. *Genome Research* 2007, **17**(12):1898-1908.

225. Schneider TD, Stephens RM: **Sequence logos: a new way to display consensus sequences**. *Nucleic Acids Res* 1990, **18**(20):6097-6100.

226. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: a sequence logo generator**. *Genome Res* 2004, **14**(6):1188-1190.

227. Rouault TA: **The role of iron regulatory proteins in mammalian iron homeostasis and disease**. *Nat Chem Biol* 2006, **2**(8):406-414.

228. Gebauer F, Hentze MW: **Molecular mechanisms of translational control**. *Nat Rev Mol Cell Biol* 2004, **5**(10):827-835.

229. Hentze MW, Gebauer F, Preiss T: **cis-Regulatory Sequences and trans-Activating Factors**. In: *Translational Control in Biology and Medicine.* Edited by Mathews M, Sonenberg N, Hershey J. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2007: 269-295.

230. Shalev A, Blair PJ, Hoffmann SC, Hirshberg B, Peculis BA, Harlan DM: **A Proinsulin Gene Splice Variant with Increased Translation Efficiency Is Expressed in Human Pancreatic Islets**. *Endocrinology* 2002, **143**(7):2541-2547.

231. Koscielny G, Le Texier V, Gopalakrishnan C, Kumanduri V, Riethoven JJ, Nardone F, Stanley E, Fallsehr C, Hofmann O, Kull M *et al*: **ASTD: The Alternative Splicing and Transcript Diversity database**. *Genomics* 2009, **93**(3):213-220.

232. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing**. *Nat Genet* 2008, **40**(12):1413-1415.

233. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes**. *Nature* 2008, **456**(7221):470-476.

234. Kozak M: **Do the 5'untranslated domains of human cDNAs challenge the rules for initiation of translation (or is it vice versa)?** *Genomics* 2000, **70**(3):396-406.

235. Kozak M: **Faulty old ideas about translational regulation paved the way for current confusion about how microRNAs function**. *Gene* 2008, **423**(2):108-115.

236. Gray NK, Hentze MW: **Iron regulatory protein prevents binding of the 43S translation pre-initiation complex to ferritin and eALAS mRNAs**. *EMBO J* 1994, **13**(16):3882-3891.

237. Muckenthaler M, Gray NK, Hentze MW: **IRP-1 binding to ferritin mRNA prevents the recruitment of the small ribosomal subunit by the cap-binding complex eIF4F**. *Mol Cell* 1998, **2**(3):383-388.

238. Rodriguez A, Hilvo M, Kytomaki L, Fleming RE, Britton RS, Bacon BR, Parkkila S: **Effects of iron loading on muscle: genome-wide mRNA expression profiling in the mouse**. *BMC Genomics* 2007, **8**:379.

239. Pestova TV, Hellen CU, Shatsky IN: **Canonical eukaryotic initiation factors determine initiation of translation by internal ribosomal entry**. *Mol Cell Biol* 1996, **16**(12):6859-6869.

240. Pestova TV, Shatsky IN, Hellen CU: **Functional dissection of eukaryotic initiation factor 4F: the 4A subunit and the central domain of the 4G subunit are sufficient to mediate internal entry of 43S preinitiation complexes**. *Mol Cell Biol* 1996, **16**(12):6870-6878.

241. Jackson RJ, Hellen CU, Pestova TV: **The mechanism of eukaryotic translation initiation and principles of its regulation**. *Nat Rev Mol Cell Biol* 2010, **11**(2):113-127.

242. Sachs MS, Geballe AP: **Downstream control of upstream open reading frames**. *Genes Dev* 2006, **20**(8):915-921.

243. Kozak M: **Structural features in eukaryotic mRNAs that modulate the initiation of translation**. *J Biol Chem* 1991, **266**(30):19867-19870.

244. Suzuki Y, Ishihara D, Sasaki M, Nakagawa H, Hata H, Tsunoda T, Watanabe M, Komatsu T, Ota T, Isogai T *et al*: **Statistical analysis of the 5' untranslated region of human mRNA using "Oligo-Capped" cDNA libraries**. *Genomics* 2000, **64**(3):286-297.

245. Warnakulasuriyarachchi D, Ungureanu NH, Holcik M: **The translation of an antiapoptotic protein HIAP2 is regulated by an upstream open reading frame**. *Cell Death Differ* 2003, **10**(8):899-904.

246. Grant CM, Miller PF, Hinnebusch AG: **Sequences 5' of the first upstream open reading frame in GCN4 mRNA are required for efficient translational reinitiation**. *Nucleic Acids Res* 1995, **23**(19):3980-3988.

247. Jackson RJ, Kaminski A, Poyry TAA: **Coupled Termination-Reinitiation Events in mRNA Translation**. In: *Translational Control in Biology and Medicine.* Edited by Mathews M, Sonenberg N, Hershey J. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2007: 197-223.

248. Elroy-Stein O, Merrick WC: **Translation Initiation via Cellular IRES**. In: *Translational Control in Biology and Medicine.* Edited by Mathews M, Sonenberg N, Hershey J. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2007: 155-172.

249. Doudna JA, Sarnow P: **Translation Initiation by Viral Internal Ribosome Entry Sites**. In: *Translational Control in Biology and Medicine.* Edited by Mathews M, Sonenberg N, Hershey J. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2007: 129-153.

250. Stumpf CR, Opperman L, Wickens M: **Chapter 14. Analysis of RNA-protein interactions using a yeast three-hybrid system**. *Methods Enzymol* 2008, **449**:295-315.

251. Sawicka K, Bushell M, Spriggs KA, Willis AE: **Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein**. *Biochem Soc Trans* 2008, **36**(Pt 4):641-647.

252. Jang SK, Wimmer E: **Cap-independent translation of encephalomyocarditis virus RNA: structural elements of the internal ribosomal entry site and involvement of a cellular 57-kD RNA-binding protein**. *Genes Dev* 1990, **4**(9):1560-1572.

253. Orom UA, Nielsen FC, Lund AH: **MicroRNA-10a binds the 5'UTR of ribosomal protein mRNAs and enhances their translation**. *Mol Cell* 2008, **30**(4):460-471.

254. Wicksteed B, Uchizono Y, Alarcon C, McCuaig JF, Shalev A, Rhodes CJ: **A cis-element in the 5' untranslated region of the preproinsulin mRNA (ppIGE) is required for glucose regulation of proinsulin translation**. *Cell Metab* 2007, **5**(3):221-227.

255. Pestova TV, Lorsch JR, Hellen CU: **The Mechanism of Translation Initiation in Eukaryotes**. In: *Translational Control in Biology and Medicine.* Edited by Mathews M, Sonenberg N, Hershey J. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2007: 87-128.

256. Lazarowitz SG, Robertson HD: **Initiator regions from the small size class of reovirus messenger RNA protected by rabbit reticulocyte ribosomes**. *Journal of Biological Chemistry* 1977, **252**(21):7842-7849.

257. Mathews DH, Sabina J, Zuker M, Turner DH: **Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure**. *J Mol Biol* 1999, **288**(5):911-940.

258. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction**. *Nucleic Acids Res* 2003, **31**(13):3406-3415.

259. Koromilas AE, Lazaris-Karatzas A, Sonenberg N: **mRNAs containing extensive secondary structure in their 5' non-coding region translate efficiently in cells overexpressing initiation factor eIF-4E**. *EMBO J* 1992, **11**(11):4153-4158.

260. Reimann I, Huth A, Thiele H, Thiele B-J: **Suppression of 15-lipoxygenase synthesis by hnRNP E1 is dependent on repetitive nature of LOX mRNA 3'-UTR control element DICE**. *Journal of Molecular Biology* 2002, **315**(5):965-974.

261. Ostareck DH, Ostareck-Lederer A, Shatsky IN, Hentze MW: **Lipoxygenase mRNA Silencing in Erythroid Differentiation: The 3'UTR Regulatory Complex Controls 60S Ribosomal Subunit Joining**. *Cell* 2001, **104**(2):281-290.

262. Ostareck-Lederer A, Ostareck DH, Standart N, Thiele BJ: **Translation of 15-lipoxygenase mRNA is inhibited by a protein that binds to a repeated sequence in the 3' untranslated region**. *EMBO J* 1994, **13**(6):1476-1481.

263. Nili M, Shinde U, Rotwein P: **Soluble RGMc/hemojuvelin is a broad spectrum BMP antagonist and inhibits both BMP2- and BMP6-mediated signaling and gene expression**. *J Biol Chem* 2010.

Appendix 1

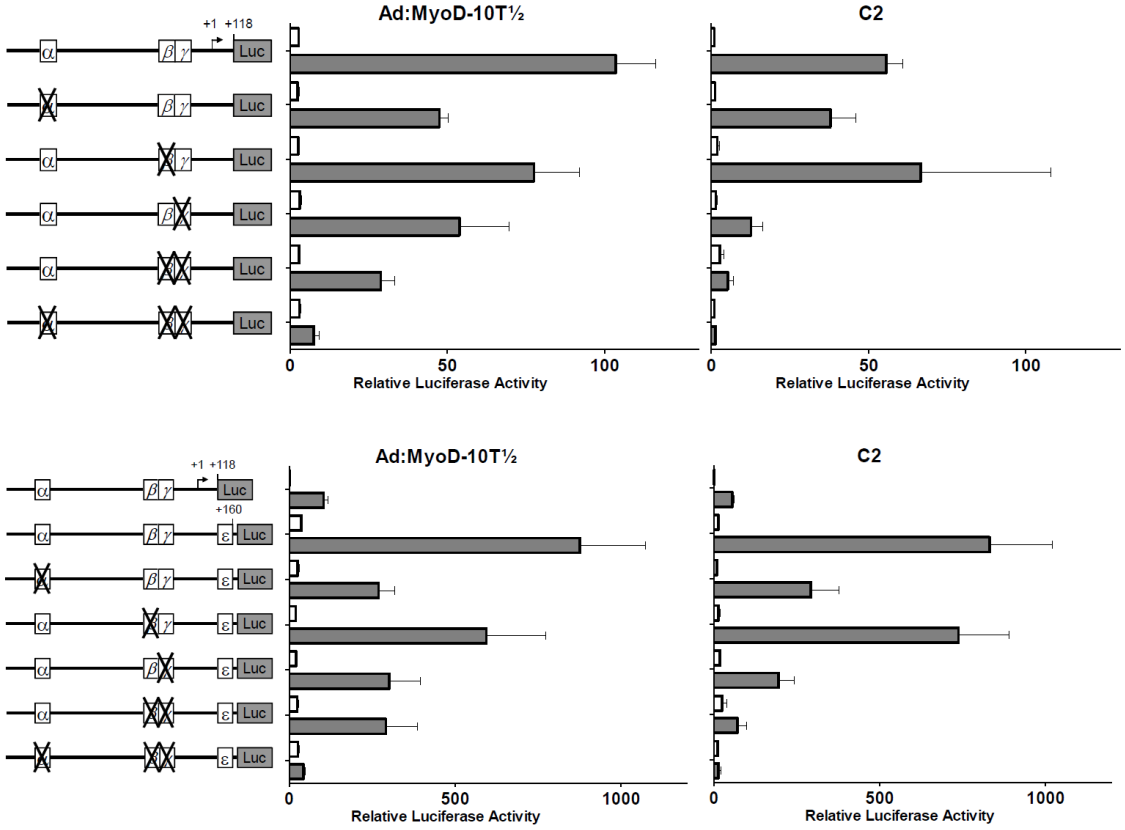


Figure A1: The ε-element does not affect the proximal promoter mutations. The Results are depicted of luciferase assays in differentiating Ad-MyoD-10T½ cells (*left panels*) and C2 myoblasts (*right panels*) transiently transfected with reporter genes containing substitution mutations of the mouse RGMc promoter (shown as an '*X*' over individual sites). In addition, the *bottom graphs* contain the ε-element. Cells were incubated in DM for 0 (*white bars*), or 24 or 48 hr (*gray bars*) before analysis. The graphs depict results of 3 - 10 independent experiments (mean ± S.E.), each performed in duplicate. See Fig. 4.1 for additional details.

Appendix 2

Curriculum Vitae

Curriculum Vitae

Christopher J. Severyn

Department of Biochemistry & Molecular Biology
School of Medicine
Oregon Health & Science University
3181 SW Sam Jackson Park Road, L-224
Portland, OR 97239

Electronic contact information:
web-site: http://openwetware.org/wiki/User:Chris
Phone: w(503) 494-0537
e-mail: severync@ohsu dot edu
fiatlux@cal.berkeley dot edu

Education:

2004-current **M.D.** Oregon Health & Science University (OHSU), Portland, Oregon, Medical Scientist Training Program (MSTP)
(anticipated completion: June 2013)

Ph.D. Biochemistry and Molecular Biology
OHSU, Portland, Oregon, *(anticipated completion: summer 2010)*
Dissertation: "Regulation and Evolutionary Origins of RGMc / Hemojuvelin expression: a muscle-enriched gene involved in iron metabolism."

1998-2002 **B.A.** Molecular & Cell Biology, University of California, Berkeley
(Emphasis: Immunology; Degree conferred Dec. 2002)

Experience:

Research:

- **Graduate Student** at OHSU (2005-2010),
  Department of Biochemistry & Molecular Biology.
  Advisor: P. Rotwein
- **Biomedical Staff Scientist** at Lawrence Livermore Nat'l Laboratory (LLNL), Biology & Biotechnology Research Division (2003-2004). BioPhysical Imaging Facility.
  Advisors: M.P. Thelen & R. Balhorn
- **Student Fellowship** at LLNL (2003).
  Advisor: M.P. Thelen
- **Internship** at LLNL, (1999) via AWU Fellowship
  Advisors: M.G. West & M.P. Thelen

Teaching:

- **Graduate Student Instructor**: OHSU Summer Undergrad Intern Program (Summer 2006)
- **Undergraduate Student Instructor** at UC Berkeley (Fall 2002 & 2003)
  Advisor: M.C. Diamond
- **Anatomy Enrichment Instructor**: Berkeley Unified School District (Spring 2001)
  Advisor: M.C. Diamond, UC Berkeley

- **Head Swimming Coach**: Sunset Swim Team (Summers 2000-02), Livermore, CA

Volunteer / Public Service:

- **OHSU Dean's Committee,** School of Medicine Strategic Plan: *Learners at all levels,* Portland, OR (2008). Chair: D.C. Dawson

- **Coordinator for Retreats and Invited Speakers, OHSU,** Portland, OR
  MSTP Retreat (2006) B.L. Davidson, U. Iowa; MSTP Springtime Keynote Address (2007) F. McCormick, UCSF; PMCB Keynote Address (2007) R. Schekman, UC Berkeley, (2010) D. Srivastava, UCSF

- **Tar-Wars Instructor**, Portland School District, OR (Spring 2005)

- **Children's Hospital Oakland**, CA, Emergency Department (2001-2004)

- **Alta Bates Ethics Committee**, Berkeley, CA**,** Community Representative (2002-2004)

- **Eagle Scout** (1997)

Honors and Awards:

- **NIH Travel Award**: Keystone Symposia on Dynamics of Eukaryotic Transcription during Development, (Apr. 2010), Big Sky, MT. Awarded by the National Institute of Environmental Health Sciences (NIEHS)

- **Ruth L. Kirschstein National Research Service Award (NRSA)** for Individual Predoctoral MD-PhD Fellows (F30), National Heart, Lung, and Blood Institute (NHLBI) (2009-2013)
  Grant number: F30-HL095327

- **OHSU Hematology/Oncology Training Program** (2007-09) Training Grant
  Grant number: T32-HL007781

- **2nd Place Oral Presentation in Musculoskeletal Research**. OHSU Student Research Forum, Portland, OR (May 2008)

- **OHSU Medical Scientist Training Program** (2004-05) Training Grant
  Grant number: T32-GM067549

- **LLNL Science & Engineering Technical Student Fellowship** (2003)

- **Associated Western Universities (AWU) Fellowship** (1999)

Professional Organizations:

- American Association for the Advancement of Science (AAAS), since 1999
- American Academy of Pediatrics, since 2009
- American Chemical Society (ACS), 2008-10
- American Medical Association (AMA), since 2004
- American Society for Biochemistry and Molecular Biology (ASBMB), since 2008
- Endocrine Society, since 2008
- Oregon Medical Association (OMA), since 2004

Meetings Attended:
- Integrating Evolution, Development, and Genomics (IEDG) 2008, Berkeley, CA
- ASBMB National Meeting April 2009, New Orleans, LA

- Keystone Symposium on Transcription during Development, April 2010, Big Sky, MT
- NHLBI Physician Scientist Trainees Conference, May 2010, Bethesda, MD

Publications (Chronological):

Peer-reviewed:

1. **C. J. Severyn**, U. Shinde, and P. Rotwein (2009) Molecular biology, genetics and biochemistry of the repulsive guidance molecule family. **Biochem J.** 422, 393-403

2. **C. Severyn** and P. Rotwein. (2010) Conserved proximal promoter elements control repulsive guidance molecule c/hemojuvelin gene transcription. **Submitted**.

3. **C. Severyn** and P. Rotwein. Repulsive Guidance Molecule C / Hemojuvelin regulation by a post-transcriptional element in the 5'-untranslated region. **In preparation June 2010.**

Abstracts:

- M.G. West, M. Hwang, S. Kadkhodayan, **C. Severyn**, M.P. Thelen (November 1999). Overexpression in hamster cells of hXRCC1 containing a debilitating mutation in the PARP-binding BRCT domain. American Society for Microbiology (ASM), DNA Repair and Mutagenesis Conference, Hilton Head, SC. UCRL-JC-135806 Abs

- **C. Severyn** and P. Rotwein (May 2008) Characterizing RGMc/Hemojuvelin Gene Expression through Functional Studies and Evolutionary Conservation. Integrating Evolution, Development, and Genomics (IEDG) Conference, Berkeley, CA.

- **C. Severyn** and P. Rotwein (April 2010) Conserved elements regulate RGMc/Hemojuvelin promoter activity in skeletal muscle. Keystone Symposium on Dynamics of Eukaryotic Transcription during Development, Big Sky, MT.

- **C. Severyn** and P. Rotwein (May 2010) Characterizing RGMc/hemojuvelin expression: a muscle-enriched gene involved in systemic iron metabolism. NHLBI Physician Scientist Trainees Conference, Bethesda, MD.

The work presented in this dissertation was performed entirely by the author, except where noted.

All the work was done between 2005-2010 in Portland, Oregon, at the